Univerzita Karlova v Praze

1. lékařská fakulta

Mechanizmy priestorového počutia a separácie zvukov

Habilitačná práca

NORBERT KOPČO

Ing., Technická Univerzita Košice, 1996

PhD., Boston University, 2003

Košice, marec 2009

Mechanizmy priestorového počutia a separácie zvukov

ABSTRAKT

Táto habilitačná práca študuje neurálne mechanizmy, ktoré človeku umožňujú využiť priestorové počutie pri spracovaní sluchových podnetov v zložitých prostrediach. Práca prezentuje behaviorálne experimenty a neurálne modely, ktoré ukazujú ako človek používa priestorový sluch pri detekovaní, identifikácií a rozpoznávaní zvukov a reči v zložitých prostrediach, napr. pri koktailových večierkoch.

Použitie priestorovej informácie pri sluchovom vnímaní závisí na zložitosti stimulov a sluchovej scény. Prvá časť práce sa zaoberá mechanizmami priestorového sluchu nachádzajúcimi sa v sluchovej periférii a v mozgovom kmeni, ktoré slúžia na spracovanie jednoduchých nerečových podnetov. Druhá časť práce je zameraná na centrálne kôrové mechanizmy priestorového vnímania, ktoré sú dôležité pre spracovanie reči v prostredí s viacerými hovoriacimi.

Prvá časť práce popisuje výsledky štyroch štúdií, ktoré skúmali, ako človek používa priestorový sluch pri detekcii nerečových zvukov maskovaných šumom. Použité cieľové zvuky boli čisté tóny, frekvenčne modulované tóny podobné vtáčiemu štebotu a amplitúdovo modulované širokospektrálne šumy. Výsledky týchto experimentov sú analyzované použitím modelov sluchovej periférie a binaurálneho spracovania zvukov v mozgovom kmeni, ktoré ukazujú, že tieto štruktúry sú rozhodujúce pri priestorovej separácii nerečových stimulov.

Druhá časť práce popisuje mechanizmy priestorového sluchu používané pri vnímaní reči v dvoch situáciách: 1) experimentálne skúma, ako závisí schopnosť človeka porozumieť hovorenej reči na priestorovej konfigurácii hovoriaceho a zdrojov rušivého nerečového zvuku; a 2) zaoberá sa mechanizmami vizuálne riadenej selektívnej pozornosti, ktoré človek používa pre zameranie sa na jedného hovoriaceho v prostredí s viacerými súbežne hovoriacimi. Výsledky týchto experimentov ukazujú, priestorová separácia reči je oveľa komplexnejšia než separácia nerečových stimulov, a to ako z hľadiska spektrálnych tak aj temporálnych aspektov. Preto tieto výsledky nie je možné popísať jednoduchými modelmi založenými na spracovaní zvuku sluchovej periférii a v podkôrových centrách.

Výsledky tejto habilitačnej práce sú príspevkom k pochopeniu neurálnych mechanizmov, ktoré umožňujú zdravému človeku robustne a presne separovať sluchové objekty a orientovať sa v zložitých akustických prostrediach. Keďže, poškodenie týchto mechanizmov vedie k dramatickému zhoršeniu sluchu v zložitých prostrediach, je ich pochopenie nevyhnutné pre ďalší vývoj metód a zariadení pre obnovu poškodeného sluchu.

Obsah

1. Úvod

Hlavnými úlohami sluchového systému človeka a zvierat je spracovanie akustických signálov a extrahovanie behaviorálne významných informácií v týchto signáloch zakódovaných (Moore, 1997). Tieto informácie môžu zahŕňať napr.: správu zakódovanú vo zvukovom signáli (lingvistický obsah reči, emočný obsah melódie), identitu zdroja signálu (hovoriaci človek, komár), a priestorovú polohu zdroja zvuku (blížiaci sa autobus). Táto habilitačná práca sa zaoberá poslednou zo zmienených funkcií: popisuje výsledky behaviorálnych experimentov a výpočtových modelov, ktoré študovali neurálne mechanizmy priestorové počutie (Blauert, 1997). Všeobecným cieľom práce je zlepšiť naše porozumenie tomu, ako ľudia určujú priestorovú polohu zdrojov zvuku, ako priestorovú informáciu používajú pri rôznych sluchových úlohách, a ako sú tieto procesy ovplyvnené štruktúrou a komplexnosťou akustického prostredia. Priestorové počutie je dôležité najmä pri 1) lokalizovaní zdroja zvuku, a 2) počúvaní zvukov maskovaných inými, rušivými zvukmi. Táto habilitačná práca je zameraná na druhú z týchto úloh: študuje použitie priestorového počutia pre separáciu zvukov, pre výber a spracovanie jedeného zo zvukov v zložitej sluchovej scéne, a mechanizmy riadiace tento výber a zameranie pozornosti na jednotlivé priestorovo odlíšené zvuky.

Priestorovému sluchu sa v poslednom storočí venovala značná pozornosť (Strutt, 1907; Gilkey and Anderson, 1997). Väčšina štúdií sa ale zameriavala na zvuky prichádzajúce zo zdrojov relatívne ďaleko od poslúcháča (nie v dosahu jeho rúk) v bezechoickom prostredí (Brungart and Durlach, 1999). Naviac, vo väčšine štúdií bola vzdialenosť zdrojov zvuku od poslucháča zafixovaná, a študovala sa len závislosť vnímania na zmene jeho horizontálnej a vertikálnej polohy (Middlebrooks and Green, 1991). Táto voľba je logická, pretože pre zdroje zvuku vo vzdialenosti väčšej ako približne jeden meter od poslucháča v bezechoickej miestnosti, sa väčšina poslucháčom používaných akustických parametrov zvuku so vzdialenosťou nemení. Toto ale neplatí pre zdroje v blízkosti poslucháča. Väčšina predkladaných štúdií sa zaoberá práve vnímaním zvukov, ktorých zdroje sú v dosahu rúk poslucháča, pre ktoré sa vnemové parametre menia aj so zmenou vzdialenosti.

Jedným z dobre známych fenoménov priestorového počutia je tzv. „efekt koktailovej párty“ (angl. "coctail party effect", Bronkhorst, 2000), ktorý popisuje

schopnosť človeka selektívne sa zamerať a spracovať informácie z jedného zdroja zvuku a ignorovať súbežné rušivé zdroje zvuku. Táto schopnosť sa u zdravo počujúcich výrazne zlepší v prípade, že sú zdroje užitočných a rušivých zvukov v priestore oddelené (Durlach and Colburn, 1978). Efekt koktailovej párty sa už v minulosti študoval pre množstvo komplexných stimulov (reč, tónové komplexy, šumové stimuly) a priestorových konfigurácií. Ale, žiadna predošlá štúdia systematicky neskúmala, ako vzdialenosť zdroja zvuku od poslucháča ovplyvňuje napr. našu schopnosť detekovať čisté tóny maskované širokospektrálnym šumom pre zdroje nachádzajúce sa v blízkosti poslucháča (t.j., pre najjednoduchší typ stimulov, pre ktorý by priestorová separácia zvukov mala viesť k zlepšeniu ich počuteľnosti). Znalosti sú ešte menej systematické pre komplexnejšie stimuly, ktoré sa môžu ľubovoľne spektrotemporálne meniť. Spektrotemporálne zmeny stimulov môžu na jednej strane poskytovať nové potenciálne zdroje informácie, ktoré môžu zlepšiť ich počutie, ale na druhej strane môžu spôsobiť, že aj keď je cieľový zvuk jasne počuteľný, nebude správne segregovaný ako cieľový zvuk, ale bude priradení k rušivému maskovaciemu zvuku (Lutfi, 1990; Oxenham et al., 2003; Arbogast and Kidd, 2000; Brungart and Simpson, 2002; Cusack et al., 2004; Alain et al., 2001). Najmenej úplné je naše porozmenie schopnostiam človeka porozumieť reči v situáciách, keď je cieľový aj maskovací zvuk rečou, napr. s podobným obsahom, polohou, alebo pohlavím hovoriacich (Durlach et al., 2003; Brungart, 2001; Bregman, 1990).

Existuje séria neurálnych modelov, ktoré popisujú ako akustickú interakciu zvuku s hlavou, torzom, a stenami miestnosťou (Shinn-Cunningham et al., 2001), tak aj neurálne spracovanie zvuku na rôznych úrovniach sluchovej dráhy (Delgutte, 1996; Hawkins and McMullen, 1996; Colburn, 1996) a kognitívne faktory ovplyvňujúce perceptuálnu organizáciu sluchovej scény (Mellinger and Mont-Reynaud, 1996). Tieto modely ale často popisujú len jednotlivé fenomény za veľmi špecifických podmienok, a je veľmi ťažké ich skombinovať za účelom popisu správania sa v prípade, že sa v scéne mení viacero parametrov naraz.

Táto habilitačná práca prezentuje sériu štúdii priestorového sluchu, ktoré kombinujú behaviorálne experimenty vykonané na ľudských subjektoch s výpočtovým neurálnym modelovaním. Spoločným cieľom týchto štúdií je poprozumenie neurálnym mechanizmom, ktoré zodpovedajú za neurálnu separáciu priestorovo oddelených zvukov a za ich spracovanie na rôznych úrovniach sluchovej dráhy.

1.1 Mechanizmy priestorového sluchu pre separáciu zvukov

Keď zaznie zvuk, napr. keď stlačíme klávesu piána, tento zvuk sa šíri od zdroja (struna) do uší poslucháča. Zvuk, ktorý dorazí do uší sa líši od zvuku, vyprodukovaného pôvodným zdrojom, a to v dôsledku interakcie zvuku s telom, hlavou, a ušnicami poslucháča (Brungart and Rabinowitz, 1999; Shinn-Cunningham et al., 2000). Naviac, ak sa poslucháč nachádza v prostredí s akusticky reflektivnými objektmi (napr. stenami), odrazy od týchto objektov sa dostávajú do uší spolu s „priamym“ zvukom. Základom priestorového počutia sú mechanizmy v sluchovom systéme poslucháča, ktoré extrahujú zo zvukov zaznamenaných ušami informácie, na základe ktorých sa potom ďalej v sluchovej dráhe vypočítava poloha zdroja zvuku. Túto informáciu poslucháči používajú napr. pri lokalizácii alebo priestorovej separácii zvukov.

1.1.1 Smerové prenosové funkcie

Transformácia, ktorou prechádza zvuk od zdroja po uši je nemenná, pokiaľ sa nemenia polohy a orientácie zdroja a poslucháča. Zdroj zvuku, prostredie v ktorom sa zvuk šíri (vrátane poslucháča a všetkých objektov a stien v prostredí) a ucho tvoria lineárny systém, ktorý transformuje vstupný signál (zvuk generovaný zdrojom) na výstupný signál (zvuk zaznamenaný uchom). Tento systém je možné matematicky charakterizovať jeho impulznou odozvou, nazývanou smerová prenosová funkcia (angl. Head-Related Transfer Function, HRTF). HRTF popisuje zvuk, ktorý sa dostane do ucha keď zdroj zvuku, nachádzajúci sa na špecifickej pozícii v okolí poslucháča, vydá širokospektrálny impulzný zvuk. Táto impulzná odozva poskytuje dostatok informácií na to, aby na jej základe bolo možné predikovať ako sa po ceste z danej polohy zdroja do ucha zmení akýkoľvek zvuk. Keďže zvuk vydaný zdrojom absolvuje cestou do každého z uší inú dráhu, pár HRTF funkcií (jedna pre ľavé a jedna pre pravé ucho) poskytuje vyčerpávajúcu informáciu o tom, aký zvuk sa dostane do uší, keď zdroj umiestnený na danej polohe vydá ľubovoľný zvuk.

Vo výskume sluchu sú dve hlavné aplikácie HRTF funkcií. Po prvé, HRTF je možné používať na generovanie virtuálneho sluchového prostredia. T.j., konvolúciou ľubovoľného zvuku so známou HRTF je možné simulovať, aký zvuk by sa dostal do uší poslucháča, ak by daný zvuk vydal zdroj na polohe zodpovedajúcej zvolenej HRTF funkcii. Po druhé, HRTF je možné analyzovať a tak určiť presne priestorové

charakteristiky zvuku, ktoré mohol sluchový systém poslucháča vyextrahovať zo zvuku, keď tento zvuk prichádzajúci z pozície zodpovedajúcej danej HRTF.

1.1.2 Akkustické charakteristiky zvuku pre priestorové počutie

Sluchový systém človeka extrahuje zo zvukov prijatých ušami dva druhy akustických priestorových charakteristík (Blauert, 1997). „Monaurálne“ charakteristiky závisia len na zvuku zaznamenanom každým uchom samostatne. „Binaurálne“ charakteristiky sú založené na porovnaní zvukov zaznamenaných oboma ušami. Najdôležitejšou monaurálnou charakteristikou je zmena v amplitúdovom spektre zvuku spôsobená interakciou medzi zvukom a hlavou, telom a ušnicou predtým, než zvuk dorazí do sluchového kanálu. Najdôležitejšími binaurálnymi charakteristikami sú rozdiely v čase príchodu (interaurálne časové rozdiely, angl. interaural time difference, ITD, ktoré je možné reprezentovať aj ako interaurálne fázové rozdiely, IPD) a rozdiely v intenzite zaznamenaného zvuku (interaurálne rozdiely v hlasitosti, angl. interaural level difference, ILD).

Monaurálne charakteristiky (angl. cues) poskytujú menej jednoznačnú informáciu o polohe zdroja zvuku než binaurálne charakteristiky, pretože sluchový systém musí pred ich použitím odhadnúť, aké spektrálne charakteristiky mal pôvodný zvuk vydaný zdrojom, aby mohol určiť, ktoré spektrálne zmeny boli spôsobené interakciou zvuku s telom, hlavou a ušnicou (t.j., spektrálne zmeny užitočné pre odhad polohy zdroja). Aj keď teoreticky nie je možné oddeliť len na základe prijatého zvuku pôvodné spektrum zvuku od spektrálnych zmien spôsobených priestorovými interakciami popísanými v HRTF, existuje množstvo zvukov, ktoré sú poslucháčom dobre známe. Podobne, ak sa neznámy zvuk prezentuje opakovane z viacerých polôh, sluchový systém poslucháča sa môže naučiť oddeliť spektrálne charakteristiky zodpovedajúce polohe od charakteristík pôvodného zvuku.

Na rozdiel od monaurálnych charakteristík sú binaurálne charakteristiky v podstate nezávislé na pôvodnom zvuku. Jediná nevyhnutná podmienka pre extrahovanie týchto charakteristík je, že vydaný zvuk musí byť dostatočne širokospektrálny, aby sluchový systém mohol extrahovať tieto charakteristiky na frekvenciách, na ktorých je na ne citlivý.

Binaurálne charakteristiky sú primárnymi charakteristikami pre vnímanie azimutálnej polohy zdroja zvuku. Pre nízkofrekvenčné zvuky (s frekvenciami pod približne 1.5 kHz) sa ITD mení relatívne výrazne s azimutálnou polohou zvuku, zatiaľ čo

ILD sa mení pomaly. Preto nie je prekvapivé, že vnímaná poloha nízkofrekvenčných zvukov je primárne určená ich ITD charakteristikou. U vysokofrekvenčných zvukov (nad približne 1.5 kHz) sa so zmenou azimutu mení hlavne ILD, keďže hlava vrhá „akustický tieň“, a tým stišuje zvuk prijatý uchom vzdialenejším od zdroja. Naviac, schopnosť vláken sluchového nervu kódovať časovú informáciu o príchode vysokofrekvenčných zvukov sa stráca, pretože tieto neuróny nedokážu páliť s frekvenciou zodpovedajúcou frekvencii vysokofrekvenčných zvukov. Preto sluchový systém pri vnímaní polohy vysokofrekfenčných zvukov zohľadňuje primárne ILD (Strutt, 1907).

1.1.3 Mechanizmy priestorovej separácie zvukov

Keď sa poslucháč snaží počúvať cieľový zvuk v prítomnosti iného súčasne znejúceho zvuku (nazývaného maskovací), schopnosť poslucháča zachytiť cieľový zvuk závisí na priestorovej polohe cieľového a maskovacieho zvuku. Vo všeobecnosti je ľahšie začuť alebo rozpoznať cieľový zvuk, keď je tento priestorovo oddelený od maskovacieho zvuku (Ebata et al., 1968; Saberi et al., 1991; Good et al., 1997; Kidd et al., 1998). Tento efekt „priestorového odmaskovania“ je určený troma faktormi. Po prvé, akustický pomer vnímanej hlasitosti cieľového a maskovacieho zvuku (angl. target-to-masker energy ratio, TMR) sa v každom uchu mení, keď sa zmení relatívna poloha cieľového a maskovacieho zvuku, a to v dôsledku akustického tieňa hlavy ako aj v dôsledku zmeny vzdialenosti zdrojov zvukov od uší. Vo všeobecnosti priestorová separácia spôsobí zvýšenie TMR (a tým zlepšenie počuteľnosti cieľa) v jednom z uší a zníženie TMR v druhom. Takže poslucháčovi stačí zamerať sa na ucho so zlepšeným TMR, a počuteľnosť priestorovo oddeleného cieľa sa zlepší.

Okrem týchto jednoduchých energetických vplyvov priestorovej separácie vedie zmena polohy jedného zo zvukov k rozdielom v binaurálnych charakteristikách každého zo zvukov. Sluchový systém je schopný detekovať prítomnosť cieľa na základe porovnania binaurálnych charakteristík celkového zvuku (ktorý je zmesou cieľového a maskovacieho zvuku) a binaurálnych charakteristík maskovacieho zvuku samotného. Vo všeobecnosti prítomnosť cieľa zmení IPD celkového zvuku najviac, keď sa IPD cieľa a maskovacieho zvuku líši maximálne. Takže, detekcia prítomnosti cieľa je najľahšia keď sa IPD cieľového a maskovacieho zvuku líši o fázu π. Podobne, detekovateľnosť cieľa v situácii, keď má

cieľový aj maskovací zvuk identickú fázu je najlepšia, keď je ILD maskovacieho zvuku 0 a ILD cieľa je ∞ (Durlach and Colburn, 1978).

Posledným faktorom ovplyvňujúcim priestorovú separáciu je perceptuálna organizácia sluchovej scény (Bregman, 1990). V závislosti na tom, ako veľmi sú si cieľový a maskovací zvuk podobné, je možné, že schopnosť poslucháča vnímať cieľový zvuk nie je obmedzená jeho počuteľnosťou, ale tým, že poslucháč nie je schopný správne prisúdiť jednotlivé komponenty počutého zvuku cieľovému a maskovaciemu zvuku (vizuálnym ekvivalentom tohto problému je oddelenie „figúry" od pozadia). Tento fenomén, v sluchovej literatúre nazývaný aj „informačným maskovaním", je tiež ovplyvnený priestorovým počutím a majú naň vplyvy aj pozornostné, krosmodálne a iné faktory. Informačné faktory hrajú dôležitú úlohu hlavne, keď je poslucháč vystavený veľkej miere neurčitosti v počutých zvukoch, napr. keď počúva náhodné komplexy tónov (Kidd et al., 1998) alebo reč jedného hovoriaceho prekrytú rečou iného hovoriaceho s podobným hlasom (Freyman et al., 1999; Hawley et al., 1999).

1.1.4 Predošlé štúdie priestorovej separácie zvukov

Priestorová separácia sa zvyčajne študuje jednou z dvoch techník: použitím slúchadiel alebo vo voľnom prostredí. Dáta zo štúdií používajúcich slúchadlá sú zamerané na rôzne špecifické aspekty spracovania sluchovej informácie (Durlach and Colburn, 1978; van de Par and Kohlrausch, 1999) a existuje niekoľko neurálnych modelov, ktoré ich úspešne popisujú (Colburn and Durlach, 1978; Colburn, 1996). Štúdie vo voľnom prostredí sa väčšinou zameriavali na relatívny príspevok energetických, binaurálnych a informačných faktorov k separácii zvukov, ako aj na aspekty vnímania, pre ktoré môže byť dôležité, že poslucháč počuje zvuk zo skutočného, nie zo simulovaného zdroja

Existuje niekoľko štúdií priestorovej separácie čistých tónov (Ebata et al., 1968; Gatehouse, 1987; Santon, 1987; Doll et al., 1992; Doll and Hanna, 1995). Tieto štúdie používali rôzne frekvencie (200 – 6000 Hz) a ukázali, že zlepšenie počuteľnosti priestorovo oddelených zvukov môže byť až 24 dB. Keď sa cieľový zvuk nahradil sériou kliknutí, priestorový zisk sa mierne znížil na 20 dB (Saberi et al., 1991; Good et al., 1997). Príspevok informačného maskovania pre komplexné zvuky môže byť až 30 dB (Watson et al., 1976; Kidd et al., 1994), čo dokazuje, že priestorová separácia môže byť veľmi

dôležitá práve v situáciách, keď je inak určenie cieľa nejednoznačné alebo veľmi ťažké. Doteraz neboli implementované žiadne kvantitatívne modely pre popis týchto výsledkov.

Väčšina štúdií priestorovej separácie bola zameraná na zmenu v zrozumiteľnosti počutej reči v závislosti na priestorovej konfigurácii sluchovej scény (Cherry, 1953). Tieto štúdie ukázali (pre prehľad pozri Bronkhorst, 2000), že príspevok binaurálneho počutia k priestorovej separácii reči maskovanej šumom je relatívne malý (okolo 3 dB). Tento výsledok sa dá čiastočne vysvetliť tým, že pre porozumenie reči je najvýznamnejšie spektrum okolo 2-5 kHz, ktoré sa neprekrýva so spektrálnou oblasťou, v ktorej je najväčší zisk separácie zvukov pre čisté tóny (100-1000 Hz). Na druhej strane, niekoľko nedávnych štúdii ukázalo, že informačné faktory hrajú významnú úlohu, obzvlášť, keď je cieľová reč maskovaná inou podobnou rečou (Hawley et al., 1999; Freyman, Balakrishnan and Helfer, 2000).

1.1.5 Výpočtové modely mechanizmov priestorového počutia

Existuje niekoľko modelov, ktoré úspešne popisujú binaurálne a priestorové počutie pre detekciu tónov (Colburn and Durlach, 1978). Teória ekvalizácie a kancelácie, tzv. E-C model (Durlach, 1972), je fenomenologický model, ktorý popisuje detekovanie zvuku ako proces, ktorý najprv časovo a hlasitostne zarovná signály počuté ľavým a pravým uchom, a potom ich navzájom odčíta (čím sa má dosiahnuť potlačenie šumu). Colburn navrhol fyziologicky plauzibilný model pre vysvetlenie týchto fenoménov, založený na znalostiach o reprezentácii zvuku v sluchovom nerve a v mozgovom kmeni (Colburn, 1973, 1977b, 1977a; Stern and Colburn, 1978; Colburn and Latimer, 1978). Extrakciou a kódovaním binaurálnych charakteristík ITD a ILD sa zaoberá aj niekoľko ďalších modelov (Marsalek, 2001; Marsalek and Kofranek, 2005).

Existuje niekoľko teórií o mechanizmoch, ktoré sluchový systém používa na detekovanie časových zmien (t.j., amplitúdovej modulácie). Najzákladnejší model predpokladá, že všetky aspekty temporálneho spracovania zvuku je možné popísať dolnopriepustným filtrom (Viemeister, 1979). Novšie modely vychádzajú z predpokladu, že na úrovni Colliculu Inferior existuje sústava modulačných filtrov, z ktorých každý je selektívny pre inú modulačnú frekvenciu (Dau, 1996). Určiť správny mechanizmus je ale relatívne zložité, keďže 1) poslucháči môžu pri počúvaní používať rôzne stratégie (napr., môžu použiť modulačnú obálku na zvolenie okamihu, kedy sa zamerajú na počutie

cieľového zvuku; Buss et al., 2003) a 2) existuje viacero jadier v sluchovej dráhe, ktoré sú citlivé na zmeny v modulácii (Joris et al., 2004).

Houtgast a jeho kolegovia navrhli niekoľko modelov pre popis zrozumiteľnosti reči v zašumenom prostredí (Plomp et al., 1980; Houtgast et al., 1980). Na ich základe boli definované dva štandardy, Artikulačný index AI (ANSI, 1969) a index zrozumiteľnosti reči, angl. Speech intelligibility index, SII (ANSI, 1997), ktorý je rozšírením AI prihliadajúcim na zmeny v amplitúdovej modulácii reči.

Zurek (1993) rozšíril Colburnov (1977) model detekcie tónov maskovaných šumom tak, aby model bolo možné použiť na predikovanie zrozumiteľnosti šumom maskovanej reči. Zurekov spôsob rozšírenia Colburnovho modelu je relatívne univerzálny, takže je možné použiť ho aj na rozšírenie iných modelov (napr. Durlachovho E-C modelu).

Zatiaľčo význam priestoru pre vizuálnu pozornosť je už podrobne preskúmaný (Desimone and Duncan, 1995; Yantis, 2005), v sluchovej doméne je toto porozumenie ešte len v základoch (Shinn-Cunningham, 2008). Existuje niekoľko štandardných modelov selektívnej sluchovej pozornosti (Treisman and Davies, 1973; Broadbent, 1958). Tieto modely popisujú pozornosť ako lokálny filter s priestorovým ohniskom a konečným priestorovým rozsahom. Ďalšie modely sú založené na mapách „saliencie“ (Kayser et al., 2005) a na neurálnych osciláciách (Wrigley and Brown, 2001). Všetky tieto modely sú ale silne hypotetické, keďže behaviorálne a neurálne charakteristiky sluchovej priestorovej pozornosti sa v súčasnosti ešte len skúmajú (Spence and Driver, 1994; Sach et al., 2000; Carlyon et al., 2001; Cusack et al., 2004; Best et al., 2007; Ebata, 2003).

1.2 Priestorová separácia a detekcia nerečových zvukov

Prvá časť tejto habilitačnej práce sa zaoberá mechanizmami priestorového sluchu pre separáciu nerečových nerečových stimulov. Sú v nej prezentované výsledky štyroch experimentálnych a modelárskych štúdií, ktoré skúmali príspevok priestorovej separácie k zlepšeniu detekovateľnosti cieľového zvuku maskovaného širokospektrálnym šumom. Všetky tieto štúdie skúmali správanie sa v najjednoduchšej sluchovej scéne pozostávajúcej len z jedného cieľového zvuku a jedného maskovacieho zvuku, a merali ako sa detekovateľnosť cieľového zvuku mení v závislosti na priestorovej konfigurácii cieľového a maskovacieho zvuku.

Všetky experimenty popísané v tejto časti práce boli vykonané na normálne počujúcich ľudských subjektoch využitím techniky simulovaného virtuálneho sluchového prostredia (Carlile, 1996). Táto technika je výhodná, pretože umožňuje presne definovať stimuly, ktoré poslucháč počuje, a tým aj presnejšie modelovať dosiahnuté výsledky. Nevýhodou techniky je obmedzená vernosť virtuálneho sluchového priestoru, ktorá môže spôsobiť, že vnímané polohy zvukov nezodpovedajú presne vnemu, ktorý by subjekty mali v skutočnom prostredí.

1.2.1 Detekovanie čistých tónov maskovaných širokospektrálnym šumom

Kapitola 2 (Kopco and Shinn-Cunningham, 2003) prezentuje výsledky štúdie, ktorá merala prahy počuteľnosti 500-Hz a 1000-Hz tónov maskovaných širokospektrálnym šumom pre rôzne priestorové konfigurácie zdrojov cieľového a maskovacieho zvuku. Všetky zdroje sa nachádzali v horizontálnej rovine v dosahu rúk subjektu (vzdialenosť zdrojov od stredu hlavy bola 15 cm alebo 1 m). Azimutálna poloha zdrojov sa menila v 45° intervaloch od -90 po +90° vo frontálnej hemisfére (t.j., pred subjektom). Tieto polohy boli simulované použitím individuálne meraných smerových prenosových funkcií HRTF.

Pre testované priestorové konfigurácie sa prah počuteľnosti menil v rozsahu až 50 dB, hlavne v dôsledku zmien v pomere intenzít cieľového a maskovacieho zvuku (angl. Target-to-Masker energy Ratio, TMR), vyplývajúcich zo zmeny ich priestorovej polohy. Práca ukázala značné rozdiely medzi subjektmi ako v individuálnych smerových prenosových funkciách HRTF, tak aj v individuálnej senzitivite ich binaurálneho sluchového systému. Kvalitatívne ale bola závislosť prahov na priestorovej konfigurácii pre všetky subjekty rovnaká. V súlade s očakávaniami vo všeobecnosti platilo, že prahy počuteľnosti sa znížili (t.j., počuteľnosť sa zlepšila) keď sa cieľový zvuk oddelil od maskovacieho zvuku v azimute. Ale v niektorých prípadoch viedla priestorová separácia zvukov k malým zmenám v počuteľnosti, alebo aj k jej miernemu zvýšeniu detekčných prahov. Veľké rozdiely medzi subjektmi boli spôsobené ako rozdielmi v monaurálnych a binaurálnych akustických charakteristikách zvukov (určenými analýzou individuálnych HRTF), tak aj individuálnymi rozdielmi vo veľkosti príspevku spracovania zvuku binaurálnych neurálnych obvodoch.

Na popis výsledkov bol implementovaný model binaurálnych interakcií v mozgovom kmeni založený na stochastickom popise aktivácie sluchových nervových vláken (Colburn, 1977a). Tento model predpokladá, že človek pri úlohe detekovať maskované tóny robí optimálne rozhodnutia založené na sub-optimálnej (zašumenej) neurálnej reprezentácii sluchového priestoru, pozostávajúcej s detektorov „koincidencie" zvukov prichádzajúcich z ľavého a pravého ucha. Predikcie tohto modelu zachytili všeobecné trendy priestorového odmaskovania v dátach. Ale, predikcie vygenerované zvlášť pre jednotlivé subjekty neboli schopné zachytiť individuálne rozdiely v počuteľnosti, a to ani po zohľadnení individuálnych rozdielov v smerových prenosových funkciách a v celkovej citlivosti binaurálnych nervových štruktúr. Tieto výsledky ukázali, že jednotliví poslucháči sa nelíšia len tým, ako sú celkovo citliví na binaurálne rozdiely v počutých zvukoch, ale že sa líšia aj v špecifickej závislosti binaurálnej citlivosti na priestorovej polohe a interaurálnych rozdieloch v maskovacom zvuku.

1.2.2 Detekovanie širokospektrálnych zvukov

Cieľmi štúdie popísanej v Kapitole 3 (Lane et al., 2004) bolo 1) skúmať príspevok priestorovej separácie k detekovateľnosti širokospektrálnych zvukov, ako aj 2) priamo porovnať výsledky psychofyzikálnych meraní na človeku s elektrofyziologickým meraním aktivácie priestorovo senzitívnych neurónov v Collicule Inferior (IC) uspanej mačky. Aby sa docielila porovnateľnosť výsledkov s výsledkami štúdie popísanej v Kapitole 2, v tejto štúdii sa urobilo niekoľko minimálnych zmien s cieľom nájsť experimentálnu paradigmu, ktorá umožní priame porovnanie ľudských behaviorálnych dát, mačacích neurálnych dát a predpovedí výpočtových modelov. Ako cieľový zvuk bol použitý 40-Hz „štebot", t.j., čistý tón, ktorého frekvencia sa cyklicky menila od 300 Hz po 1.5 alebo 12 kHz s periódou 12.5 ms tak, že dlhodobá spektrálna obálka stimulu bola konštantná.

Hlavným zámerom tejto štúdie bolo skúmať, či sa príspevok priestorovej separácie k odmaskovaniu zmení pre širokospektrálne zvuky, ktoré sú spracované viacerými periférnymi kanálmi, v porovnaní s čistými tónmi, ktoré sú spracované primárne jedným z periférnych kanálov. Naviac sa predpokladalo, že tak ako v predošlej štúdii, aj tu budú výsledky ovplyvnené priestorovými zmenami v TMR ako aj binaurálnym spracovaním informácie v mozgovom kmeni. Keďže vysokofrekvenčné zvuky sú v porovnaní s nízkofrekvenčnými zvukmi silnejšie ovplyvnené akustickým tieňom vrhaným hlavou (a

tým sa u nich viac prejaví efekt TMR), zatiaľ čo binaurálne neurálne obvody sú relatívne necitlivé k zmenám v presnom časovaní vysokofrekvenčných zvukov (a tým by u nich binaurálny príspevok mal byť malý). Na rozdiel od prvej štúdie boli v tieto štúdie v prípade ľudí aj mačiek vykonané vo virtuálnom sluchovom prostredí vytvorenom použitím neindividualizovaných HRTF, pričom sa už nemenila ani vzdialenosť zvukov a meral sa len prah počuteľnosti v závislosti na horizontálnej polohe cieľového a maskovacieho zvuku.

Výsledky psychofyzikálnej štúdie ukázali, že prahy citlivosti boli podobné pre širokospektrálne a hornopriepustne filtrované stimuly, ako aj pre monaurálne a binaurálne stimuly. V protiklade, prahy pre dolnopriepustne filtrované stimuly boli horšie. Tieto výsledky naznačili, že detekcia širokospektrálnych stimulov je v prvom rade ovplyvnená monaurálnymi faktormi súvisiacimi so zmenami TMR v uchu, v ktorom priestorová separácia cieľového a maskovacieho zvuku vedie k zvýšeniu TMR. V súlade s tým, väčšina neurónov v IC mačky citlivých na vysoké frekvencie mala neurálny prah detekovateľnosti určený zmenou TMR v jednom z uší.

Na druhej strane, psychofyzikálne výsledky pre nízkofrekvenčné stimuly záviseli rovnako na TMR ako aj na binaurálnych charakteristikách zvuku. Podobne, nízkofrekvenčné neuróny v IC citlivé na ITD vykazovali zmeny v prahoch v závislosti na polohe maskovacieho šumu, konzistentné s modelmi binaurálnych jadier založenými na kroskorelácii. Tento trend bol obzvlášť výrazný, ak sa sledovala zmena citlivosti prahu celej populácie neurónov.

Psychofyzikálne dáta boli relatívne dobre predikované populačným modelom, ktorý uvažoval, že prah počuteľnosti širokospektrálneho zvuku je možné určiť nájdením periférneho kanálu, v ktorom je po separovaní zdrojov výsledné TMR najpriaznivejšie. Kvalitatívne sa predikcie tohto modelu podobali predikciám kroskorelačného modelu popisujúceho fyziologické dáta. Tieto dva modely sa ale líšili v jednom dôležitom aspekte: zatiaľ čo model psychofyzikálnych dát predpokladal, že spracovanie stimulov je čiste monaurálne, model fyziologických dát bol založený na aktivite binaurálne citlivých neurónov. Preto sa na základe štúdie nedá jednoznačne určiť, aký je neurálny substrát, ktorý je za tieto výsledky zodpovedný, monaurálny alebo binaurálny.

1.2.3 Integrácia informácií cez frekvencie pri priestorovej separácii širokospektrálnych zvukov

Jedným z neočakávaných výsledkov ľudského psychofyzikálneho experimentu z predošlej štúdie bolo, že detekovateľnosť vysokofrekvenčných (t.j., hornopriepustne filtrovaných) a širokospektrálnych zvukov bola vždy lepšia než počuteľnosť nízkofrekvenčných (t.j., dolnopriepustne filtrovaných) zvukov (Lane et al., 2004). Tento výsledok bol ale v protiklade s predpokladom, že nízkofrekvenčné zvuky budú detekovateľné lepšie, pretože pre ne je prah počuteľnosti výsledkom kombinácie binaurálneho a monaurálneho (TMR) spracovania, zatiaľ čo pre vysokofrekvenčné zvuky je prah určený len monaurálnym spracovaním. Kapitola 4 (Kopco, 2005) popisuje výsledky experimentu, ktorý testoval dve hypotézy týkajúce sa možného dôvodu tejto nekonzistentnosti. Jedna hypotéza bola, že existuje silná integrácia informácie z periférnych kanálov, ktorá spôsobuje, že širokospektrálne (a vysokofrekvenčné) prahy sú lepšie než nízkofrekvenčné prahy, napriek príspevku binaurálnej informácie pre nízkofrekvenčné stimuly. Druhá hypotéza vychádzala z možnosti, že je nesprávny niektorý z predpokladov, na ktorých bol založený model používaný pri popise predošlých psychofyzikálnych dát. Špecificky, tento model predpokladal, že prah detekovateľnosti vyjadrený ako TMR je nezávislý na frekvencii stimulu. Ak by tento predpoklad nebol správny, a ak by sa prah s frekvenciou periférneho kanálu zlepšoval, výsledkom by bol rozdiel podobný tomu, ktorý bol pozorovaný v predošlej štúdii. Na otestovanie týchto dvoch hypotéz sa vykonala nová psychofyzikálna štúdia, ktorá používala podobné metódy ako Lane (2004). V tejto novej štúdii sa ale okrem pôvodných zvukov merali prahy aj pre úzkospektrálne zvuky získané prefiltrovaním pôvodných zvukov cez model periférneho kanálu s najpriaznivejším pomerom TMR (čím sa umožnilo priame otestovanie príspevku integrácie cez frekvencie) ako aj pre monaurálne prezentované zvuky (čím sa umožnilo priame testovanie binaurálnych príspevkov). Štúdia našla veľké rozdiely (až 10 dB) medzi prahmi v jednotlivých meraniach, pričom binaurálne prahy boli vždy lepšie než zodpovedajúce monaurálne prahy, ktoré boli zase lepšie než zodpovedajúce predfiltrované jednokanálové prahy. Porovnanie jednokanálových prahov pre kanály s rozdielnou stredovou frekvenciou ale jednoznačne ukázalo, že detekcia je lepšia ak je kanál na vysokej frekvencii. Tento výsledok potvrdil druhú hypotézu a priamo ukázal, že detekčný

prah (vyjadrený ako prahový pomer TMR) je závislý na frekvencii sledovaného periférneho kanálu, a že model, ktorý by dokázal popísať tieto dáta presne, musí zahrnúť binaurálne spracovanie, ale nie kombinovanie informácie cez nezávislé periférne kanály.

1.2.4 Perceptuálne kombinovanie informácie o časovej modulácii a priestorovej separácii

Aj keď predchádzajúce dve štúdie boli primárne zamerané na určenie efektu priestorovej separácie na zlepšenie počutia, stimuly v nich použité sa menili aj v čase. Preto je možné, že ich výsledky boli ovplyvnené aj interakciami medzi detektormi priestorových a časových charakteristík na neurálnej úrovni (neuróny v IC sú citlivé na priestorovú polohu aj na amplitúdovú moduláciu) alebo na kognitívnej úrovni (subjekty mohli zamerať svoju pozornosť na „vyhľadávanie“ časových zmien v počutom stimule alebo na „vyhľadávanie“ priestorovo separovaného cieľa, alebo na kombinovnie oboch informácií). Cieľom štúdie popísanej v Kapitole 5 (Kopco and Shinn-Cunningham, 2008) bolo psychofyzikálne určiť, ako poslucháči kombinujú priestorové a modulačné informácie pri detekcii maskovaných zvukov. Aby sa minimalizovala možnosť, že subjekty použijú pri tejto úlohe iné informácie ako tie, na ktoré je štúdia zameraná, v tejto štúdii sa použil širokospektrálny šum ako cieľový aj maskovací zvuk (pričom šumy použité ako cieľový a maskovací šum boli navzájom nezávislé). Čiže cieľový a maskovací zvuk sa líšili len v želaných dvoch aspektoch. Vyšetrovali sa tri hlavné hypotézy:

> H1. Kombinovaný efekt priestorovej separácie a prítomnosti modulácie bude asymetrický. T.j., zlepšenie počuteľnosti pri prítomnosti oboch informácií bude závisieť na tom, či je modulácia prítomná v cieľovom zvuku, v maskovacom zvuku, alebo v oboch zvukoch (a nie len na tom, či cieľový a maskovací zvuk majú odlišnú moduláciu).
>
> H2. Efekt modulácie na priestorové odmaskovanie bude závisieť na konkrétnych polohách, na ktorých sa budú separované zdroje cieľového a maskovacieho zvuku nachádzať (t.j., nie len na tom, či sú umiestnené na tom istom alebo na rozdielnom mieste). Túto závislosť je možné očakávať, ak spôsob, ktorým ľudia kombinujú priestorovú a modulačnú informáciu, je založený na subkortikálnych priestorových reprezentáciách, analyzovaných v predošlých kapitolách. Ak je ale založený na centrálnejšej, viac abstraktnej reprezentácii

priestoru (necitlivej napr. na frekvenciu stimulov), potom by sa táto hypotéza nemala potvrdiť.

H3. Kombinovaný príspevok priestorovej a modulačnej informácie nebude aditívny. Ak by sa informácie o modulácii a priestorovej separácii spracovávali nezávislo a ich kombinácia by bola optimálna, potom zlepšenie počutia modulovaných priestorovo oddelených zvukov by malo byť predikované ako súčet zlepšení pozorovaných, keď sú zvuky len oddelené (ale nie rozdielne modulované) a keď sú zvuky rozdielne modulované (ale nie priestorovo separované). Na druhej strane, ak pri kombinovaní informácii hrá dôležitú úlohu perceptuálna organizácia, potom je možné, že zisk z poskytnutia oboch informácii súčasne bude väčší ako súčet jednotlivých príspevkov; resp., ak sa ľudia vždy sústredia len na jeden z aspektov, potom zisk z poskytnutia oboch informácii bude menší ako súčet jednotlivých príspevkov.

Štúdia ukázala, že kombinovaný efekt je nesymetrický z hľadiska prítomnosti modulácie v cieľovom alebo maskovacom zvuku, potvrdzujúc hypotézu H1. Taktiež štúdia ukázala, že závislosť prahov počuteľnosti na type modulácie je menšia, keď sú zvuky priestorovo oddelené, než keď sú na tom istom mieste. To znamená, že kombinovanie priestorovej a modulačnej informácie je subaditívne, potvrdzujúcu hypotézu H3. Na druhej strane, vyvracajúc hypotézu H2, štúdia nenašla žiadnu štatistickú interakciu medzi efektom modulácie a konkrétnymi polohami, na ktorých sa priestorovo oddelené stimuly nachádzali. Tento výsledok naznačuje, že kombinovanie modulačnej a priestorovej informácie sa deje na kortikálnej úrovni, na ktorej priestor už nie je reprezentovaný binaurálnymi charakteristikami.

Na popis efektu modulácie na priestorové odmaskovanie boli navrhnuté dva výpočtové modely. Jeden predpokladal, že v IC existujú špecifické neurálne detektory modulácie, a že poslucháč používa informáciu z týchto detektorov na identifikovanie zmien v hĺbke modulácie v celkovom stimule. Druhý model predpokladal, že poslucháč sleduje obálku maskovacieho zvuku, a že svoje rozhodnutia zakladá na detekovaní zmeny celkovej intenzity zvuku v časových okamihoch, kedy je pomer TMR najpriaznivejší. Výsledky boli viac konzistentné s prvým modelom, potvrdzujúc, že poslucháči pri svojom rozhodovaní používajú subaditívnu kombináciu modulačnej a priestorovej informácie.

1.3 Priestorová separácia s porozumenie hovorenej reči

Predchádzajúce štúdie (Kapitoly 2-5) ukázali, aké zložité je spracovanie priestorovej informácie pri separácii zvukov už pri tej najjednoduchšej úlohe (detekcia prítomnosti zvuku) a pre relatívne jednoduché stimuly (signály s jednoducho popísanou spektrotemporálnou štruktúrou). Štúdium vnímania týchto stimulov je nevyhnutným krokom na ceste k popisu toho, ako ľudia používajú priestor na separáciu a porozumenie hovorenej reči, ktorá má oveľa zložitejšiu a premenlivejšiu spektrotemporálnu štruktúru (nehovoriac o jej lingvistických, krosmodálnych, a iných kontextuálnych aspektoch).

V tejto časti práce sú prezentované dve štúdie. Prvá skúmala ako človek používa priestor pre separáciu reči maskovanej šumom (Kapitola 6) a druhá ako človek presúva svoju pozornosť z jedného miesta (objektu) na druhé pri počúvaní jedného z viacerých súbežne hovoriacich (Kapitola 7).

1.3.1 Priestorová separácia a porozumenie reči maskovanej šumom

Príspevok priestorovej separácie k porozumeniu reči maskovanej šumom sa tradične študoval v priestorových konfiguráciách, v ktorých cieľové aj maskovacie zvuky boli rovnako vzdialené aspoň 2 metre od poslucháča. Štúdia prezentovaná v Kapitole 6 (Schickler et al., 2000) skúmala tento efekt pre konfigurácie, pri ktorých sa poloha simulovaných zdrojov zvuku menila v azimute pričom zvuky mohli byť blízko alebo ďaleko od hlavy. Cieľovými zvukmi bola reč (gramaticky správne ale sémanticky nezmyselné vety v severoamerickej angličtine, napr. „The right cane guards an edge.“) nahovorená mužskými hlasmi. Maskovacím zvukom bol náhodný šum s dlhodobým spektrom zhodným s priemernou obálkou reči používanej ako cieľové zvuky. Štúdia bola vykonaná v simulovanom virtuálnsom sluchovom prostredí (t.j., stimuly boli prezentované cez slúchadlá použitím HRTF odvodených zo sférického modelu hlavy). Príspevok priestorového odmaskovania sa meral použitím adaptívnej metódy, pri ktorej bola zafixovaná hlasitosť maskovacieho šumu (v uchu v ktorom bol pomer TMR priaznivejší) a adaptívne sa menila hlasitosť prezentovaných viet. Výsledný prah počuteľnosti zodpovedal TMR, pri ktorom subjekt správne identifikoval 50% prezentovaných slov.

Štúdia ukázala, že malé zmeny v polohe hovoriaceho a/alebo zdroja maskovacieho zvuku môžu viesť k veľkým zmenám v zrozumiteľnosti, keď sa zdroje zvukov nachádzajú v blízkosti poslucháča. Táto citlivosť je dôsledkom toho, že v blízkosti poslucháča aj malé

zmeny v polohe hovoriaceho vedú k veľkým zmenám v celkovej hlasitosti počutých zvukov. Naviac, keďže zvuky prichádzajúce z blízkosti uší majú veľké interaurálne rozdiely v hlasitosti, výrazne sa mení aj príspevok binaurálnych a priestorových mozgových analýz k potlačeniu šumu odmaskovaniu. Na popis výsledkov bol použitý existujúci fenomenologický model priestorového príspevku k porozumeniu reči v zašumenom prostredí (Zurek, 1993). Tento model vychádza z Colburnovho modelu spracovania zvuku v mozgovom kmeni (Kapitola 2), a predpokladá, že príspevok jednotlivých frekvenčných kanálov k porozumeniu reči je priamo úmerný ich príspevku k zlepšenej detekovateľnosti tónov, pričom jednotlivé frekvenčné kanály sú prevážené priemernou informačnou hodnotou zodpovedajúcou významu daného kanálu pre porozumenie reči. Predikcie tohto modelu binaurálnej zrozumiteľnosti reči dobre popísali výsledky pre priestorové konfigurácie, ktoré sa testovali už v predošlých štúdiách. Vo zvyšných priestorových konfiguráciách sa ukázali malé ale dôležité rozdiely medzi predikciami modelu a nameranými prahmi, obzvlášť ak sa model použil na predikovanie percenta správne porozumených slov (nie TMR prahu zodpovedajúcemu 50% zrozumiteľnosti). Tieto výsledky naznačujú, že súčasné teórie nie sú schopné presne popísať vplyv priestorovej separácie na zrozumiteľnosť reči v niektorých novo skúmaných konfiguráciách.

1.3.2 Selektívna pozornosť pri porozumení reči jedného z viacerých hovoriacich

Veľmi bežná, ale zároveň veľmi zložitá, je sluchová scéna, v ktorej je viacero súbežne hovoriacich a poslucháč sa snaží zamerať svoju pozornosť na jedného z nich. V takejto situácii môže poslucháč identifikovať reč počúvaného hovoriaceho nesprávne nie len preto, že táto reč je zamaskovaná („prekričaná“) inými hovoriacimi, ale aj preto, že si môže pomýliť, ktorého hovoriaceho práve počul, alebo kde sa hovoriaci, ktorého práve počúva, nachádza. Situácia sa stáva ešte zložitejšou v prípade, že sa poloha a/alebo hlas počúvaného hovoriaceho dynamicky mení. Cieľom poslednej štúdie, popísanej v Kapitole 7 (Best et al., 2008), bolo skúmať faktory, ktoré ovplyvňujú schopnosť človeka dynamicky presúvať priestorovú sluchovú pozornosť v prostredí s viacerými súčasne hovoriacimi.

V štúdii sedel subjekt pred piatimi reproduktormi rozmiestnenými v štvrťkruhu pred ním, a s jednou svietivou diódou (LED) umiestnenou na každom z reproduktorov. Pri jednom meraní zaznela súčasne z každého z reproduktorov séria štyroch čísel, pričom na

každý reproduktor pripadalo v každom časovom okamihu iné číslo vyslovené iným hovoriacim. Počas prezentácie rečového stimulu sa zároveň na jednotlivých reproduktoroch rozsvecovali LEDky (pre každý zo štyroch časových krokov jedna). Úlohou poslucháča bolo po počutí stimulu na klávesnici zadať sériu čísel počutú z reproduktorov, na ktorých sa vysvietili LEDky. Náročnosť úlohy sa menila vsúvaním tichých intervalov v rozsahu 0 až 1000 ms medzi jednotlivé časové kroky. Dlhší interval znamenal, že subjekt mal viac času na zanalyzovanie práve počutých čísel, ako aj na preorientovanie svojej pozornosti na potenciálne novú polohu, z ktorej príde cieľové číslo v nasledujúcom kroku. Identifikácia čísel sa merala v troch typoch scén. V prvom type sa poloha cieľového reproduktora behom jednej prezentácie nemenila. V druhom prípade sa menila s tým, že LEDka určujúca, kam má poslucháč svoju pozornosť zamerať v ďalšom kroku sa rozsvietila až ukončení tichej pauzy, t.j., synchrónne so začiatkom nového slova. V treťom prípade sa poloha tiež menila, ale LEDka určujúca polohu cieľového reproduktora sa rozsvietila vždy už na začiatku tichej pauzy, čím sa poslucháčovi umožnilo začať svoju pozornosť presúvať už počas pauzy, a očakávať príchod nového slova zo zameraného reproduktora. Vykonali sa dva experimenty, v jednom sa cieľový hlas behom jednej prezentácie nemenil, nezávislo od toho, ktorý typ scény sa práve použil. To znamená, že poslucháč teoreticky nemusel sledovať polohu hovoriaceho. Stačilo, ak ho bol schopný behom celej rečovej sekvencie identifikovať. V druhom experiment sa cieľový hlas náhodne menil. Takže poslucháč musel zamerať pozornosť len na vizuálne definovanú cieľovú pozíciu.

Štúdia ukázala, že presnosť identifikácie čísel bola najvyššia, ak človek mohol ponechať pozornosť zameranú na jednu pozíciu počas celej sekvencie, než keď sa pozornosť presúvala z jedného miesta na druhé. Stratu presnosti spôsobenú zvýšenou kognitívnou záťažou súvisiacou s presúvaním pozornosti súbjekty neboli schopné eliminovať ani keď interval medzi slovami bol celá 1 sekunda, počas ktorej sa subjekt mohol zamerať na novú polohu následného čísla. Nielen, že priestorová kontinuita eliminovala už v minulosti popísané straty presnosti identifikácie súvisiace s presúvaním priestorovej pozornosti, ale umožnila aj postupné zlepšovanie priestorovej selektivity a tým aj zlepšenú identifikáciu neskôr prichádzajúcich čísel. Ak sa naviac nemenil ani hlas cieľového hovoriaceho, toto dolaďovanie selektívnej pozornosti sa ešte zvýraznilo. Tieto výsledky ukázali, že ak sa zameranie pozornosti ponechá na jednom objekte v komplexnej

sluchovej scéne, pozornostná selektivita sa zlepšuje, a to aj v prípade, že je na tom istom mieste ponechaná po niekoľko sekúnd.

Analýza chýb, ktoré subjekty v experimente robili, ukázala, že najčastejšími chybami bolo udanie čísla, ktoré bolo prezentované zo susedného reproduktora k cieľovému. Pravdepodobnosť takýchto zámien klesala so vzdialenosťou medzi cieľovým a identifikovaným reproduktorom. Takáto distribúcia chýb je konzistentná s modelom selektívnej pozornosti nazývaným „javiskový reflektor“ (angl. spotlight; Treisman, 1971). Naviac, porovnanie tvaru distribúcie chýb pre rôzne typy merania ukázalo, že tento filter sa postupne zaostruje, obzvlášť pre cieľové stimuly s v čase konštantným hlasom a polohou.

1.4 Zhrnutie habilitačnej práce a ďalších výsledkov autora

Neurálne mechanizmy, ktoré človeku umožňujú orientovať sa v sluchovom priestore, sú zložité a v mnohých ohľadoch dosiaľ neprebádané. Táto habilitačná práca osvetľuje niekoľko aspektov použitia priestorového sluchu pre separáciu zvukov v zložitých akustických prostrediach, ako aj neurálnu bázu, na ktorej je toto správanie založené. Najdôležitejšie nové poznatky popísané v práci osvetľujú, ako človek používa priestorový sluch na separáciu nerečových a rečových stimulov, ktorých zdroje sa nachádzajú v dosahu poslucháča, a tým mu umožňujú priamu interakciu. Ďalej práca ukazuje, že obzvlášť pre nerečové stimuly, ktoré sú analyzované primárne v podkôrových mozgových oblastiach, je zhoda medzi behavioralnými experimentmi na človeku a neurálnymi dátami získanými v talame mačky veľmi blízka. Na druhej strane, mechanizmy a stratégie, ktoré človek používa na kombinovanie informácií pri vytváraní perceptuálnych objektov a pri orientovaní pozornosti v sluchovej scéne sú oveľa menej dobre preskúmané. Práca popisuje niekoľko nových poznatkov osvetľujúcich tieto mechanizmy.

Všetky výsledky popísané v tejto práci môžu byť užitočné pre pochopenie dôsledkov zhoršenia sluchu na počutie v každodenných situáciách. Naviac, pri súčasnom rozvoji technických možností prostetických zariadení a iných nových technológií sa obmedzujúcim faktorom pri snahe zlepšiť alebo obnoviť sluch pacientov stáva základné porozumenie neurálnym mechanizmom, ktoré sú zodpovedné za perceptuálne a kognitívne schopnosti u zdravých jedincov. Preto okrem čiste teoretickej stránky môžu byť výsledky

predloženej práce užitočné napr. pri vývoji nových načúvacích strojčekov, kochleárnych implantátov, alebo virtuálnych sluchových displejov pre slepcov.

Okrem výskumu zameraného na priestorovú separáciu zvukov sa autor v posledných rokoch zaoberal aj výskumom iných aspektov priestorového počúvania, ako aj štúdiom neurálne inšpirovaných počítačových algoritmov pre učenie sa a rozpoznávanie vzorov v zložitých dátach. Výsledky týchto štúdií presahujú tému tejto habilitačnej práce. Preto sú tu len krátko zhrnuté s odkazom na relevantné verejne dostupné publikácie. Časť výskumu sa venovala schopnosti človeka vnímať polohu zdroja zvuku, a to vykonaním analýzy akustických parametrov, ktoré sa v mozgu zo zvukov extrahujú pri lokalizácii (Shinn-Cunningham et al., 2000; Shinn-Cunningham et al., 2005). Ďalšie experimenty sa zaoberali efektom časovej súslednosti zvukov na schopnosť človeka lokalizovať zvuk (Kopco et al., 2007a). Posledné dve sluchové štúdie sa týkali neurálnej plasticity a učenia sa pri priestorovom sluchovom vnímaní. Jedna štúdia skúmala procesy učenia pri vnímaní vzdialenosti zdrojov zvuku (Kopco et al., 2004; Schoolmaster et al., 2003). Druhá študía skúmala „bruchomluvecký efekt“, t.j., mechanizmy vizuálne vyvolanej plasticity priestorových sluchových máp, a to u človeka aj u iných primátov (Lin et al., 2007; Kopco et al., 2007b). Relatívne nezávislou témou výskumu boli učiace sa neurálne algoritmy, kde autor vyvinul nový algoritmus pre robustnú klasifikáciu zašumených viacdimenzionálnych dát (Kopco and Carpenter, 2004).

Referencie

Alain C, Arnott SR, Picton TW (2001) Bottom-up and top-down influences on auditory scene analysis: Evidence from event-related brain potentials. Journal of Experimental Psychology: Human Perception and Performance 27:1072-1089.

ANSI (1969) American national standard methods for the calculation of the Articulation Index, S3.5. New York: American National Standards Institute, Inc.

ANSI (1997) Methods for calculation of the speech intelligibility index, ANSI S3.5-1997. New York: American National Standards Institute, Inc.

Arbogast TL, Kidd J, Gerald (2000) Evidence for spatial tuning in informational masking using the probe-signal method. Journal of the Acoustical Society of America 108:1803-1810.

Best V, Ozmeral E, Shinn-Cunningham BG (2007) Visually guided attention enhances target identification in a complex auditory scene. J Assoc Res Otolaryngol 8:294-304.

Best V, Ozmeral EJ, Kopco N, Shinn-Cunningham BG (2008) Object continuity enhances selective auditory attention. Proceedings of the National Academy of Sciences of the USA 105:13174-13178.

Blauert J (1997) Spatial Hearing, 2nd Edition. Cambridge, MA: MIT Press.

Bregman AS (1990) Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA: MIT Press.

Broadbent DE (1958) Perception and communication. Elmsford, NJ: Pergamon.

Bronkhorst AW (2000) The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions. Acustica 86:117-128.

Brungart DS (2001) Informational and energetic masking effects in the perception of two simultaneous talkers. J Acoust Soc Am 109:1101-1109.

Brungart DS, Rabinowitz WM (1999) Auditory localization of nearby sources I: Head-related transfer functions. Journal of the Acoustical Society of America 106:1465-1479.

Brungart DS, Durlach NI (1999) Auditory localization of nearby sources II: Localization of a broadband source in the near field. Journal of the Acoustical Society of America 106:1956-1968.

Brungart DS, Simpson BD (2002) The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. Journal of the Acoustical Society of America 112:664-676.

Buss E, Hall JW, III, Grose JH (2003) The masking level difference for signals placed in masker envelope minima and maxima. J Acoust Soc Am 114:1557-1564.

Carlile S (1996) Virtual Auditory Space: Generation and Applications. New York: RG Landes.

Carlyon RP, Cusack R, Foxton JM, Robertson IH (2001) Effects of attention and unilateral neglect on auditory stream segregation. J Exp Psychol Hum Percept Perform 27:115-127.

Cherry EC (1953) Some experiments on the recognition of speech, with one and with two ears. Journal of the Acoustical Society of America 25:975-979.

Colburn HS (1973) Theory of binaural interaction based on auditory-nerve data. I: General strategy and preliminary results on interaural discrimination. Journal of the Acoustical Society of America 54:1458-1470.

Colburn HS (1977a) Theory of binaural interaction based on auditory-nerve data. II: Detection of tones in noise. Journal of the Acoustical Society of America 64:525-533.

Colburn HS (1977b) Theory of binaural interaction based on auditory-nerve data. II: Detection of tones in noise. Supplementary material. Journal of the Acoustical Society of America AIP document no. PAPS JASMA-61-525-98.

Colburn HS (1996) Binaural Models. In: Auditory Computation (Hawkins HL, McMullen TA, Popper AN, Fay RR, eds), pp 332-400. New York: Springer Verlag.

Colburn HS, Latimer JS (1978) Theory of binaural interaction based on auditory-nerve data. III: Joint dependence on interaural time and amplitude differences in discrimination and detection. Journal of the Acoustical Society of America 64:96-106.

Colburn HS, Durlach NI (1978) Models of binaural interaction. In: Handbook of Perception (Carterette EC, Friedman MP, eds), pp 467-518. New York: Academic Press.

Cusack R, Deeks J, Aikman G, Carlyon RP (2004) Effects of location, frequency region, and time course of selective attention on auditory scene analysis. J Exp Psychol Hum Percept Perform 30:643-656.

Dau T (1996) Modeling auditory processing of amplitude modulation. In: Universität Oldenburg, Germany.

Delgutte B (1996) Physiological models for basic auditory percepts. In: Auditory Computation (Hawkins HL, McMullen TA, Popper AN, Fay RR, eds), pp 157-220. New York: Springer Verlag.

Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18:193-222.

Doll TJ, Hanna TE (1995) Spatial and spectral release from masking in three-dimensional auditory displays. Human Factors 37:341-355.

Doll TJ, Hanna TE, Russotti JS (1992) Masking in three-dimensional auditory displays. Human Factors 34:255-265.

Durlach NI (1972) Binaural signal detection: Equalization and cancellation theory. In: Foundations of Modern Auditory Theory (Tobias JV, ed). New York: Academic Press.

Durlach NI, Colburn HS (1978) Binaural phenomena. In: Handbook of Perception (Carterette EC, Friedman MP, eds), pp 365-466. New York: Academic Press.

Durlach NI, Mason CR, Kidd G, Jr., Arbogast TL, Colburn HS, Shinn-Cunningham BG (2003) Note on informational masking. J Acoust Soc Am 113:2984-2987.

Ebata M (2003) Spatial unmasking and attention related to the cocktail party problem. Acoustical Science and Technology 24:208-219.

Ebata M, Sone T, Nimura T (1968) Improvement of hearing ability by directional information. Journal of the Acoustical Society of America 43:289-297.

Freyman RL, Helfer KS, McCall DD, Clifton RK (1999) The role of perceived spatial separation in the unmasking of speech. Journal of the Acoustical Society of America 106:3578-3588.

Gatehouse RW (1987) Further research on free-field masking. Journal of the Acoustical Society of America 82:S108.

Gilkey R, Anderson T (1997) Binaural and Spatial Hearing in Real and Virtual Environments. Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc.

Good MD, Gilkey RH, Ball JM (1997) The relation between detection in noise and localization in noise in the free field. In: Binaural and Spatial Hearing in Real and Virtual Environments (Gilkey R, Anderson T, eds), pp 349-376. New York: Erlbaum.

Hawkins HL, McMullen TA (1996) Auditory Computation: An Overview. In: Auditory Computation (Hawkins HL, McMullen TA, Popper AN, Fay RR, eds), pp 1-14. New York: Springer Verlag.

Hawley ML, Litovsky RY, Colburn HS (1999) Speech intelligibility and localization in a multi-source environment. Journal of the Acoustical Society of America 105:3436-3448.

Houtgast T, Steeneken HJM, Plomp R (1980) Predicting speech intelligibility in rooms from the modulation transfer function I. General room acoustics. Acustica 46:60-72.

Joris PX, Schreiner CE, Rees A (2004) Neural processing of amplitude-modulated sounds. Physiol Rev 84:541-577.

Kayser C, Petkov CI, Lippert M, Logothetis NK (2005) Mechanisms for Allocating Auditory Attention: An Auditory Saliency Map. Current Biology 15:1943-1947.

Kidd G, Mason CR, Deliwala PS, Woods WS, Colburn HS (1994) Reducing informational masking by sound segregation. Journal of the Acoustical Society of America 95:3475-3480.

Kidd J, Gerald, Mason CR, Rohtla TL, Deliwala PS (1998) Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. Journal of the Acoustical Society of America 104:422-431.

Kopco N (2005) Across-frequency integration in spatial release from masking. In: Proceedings of the European Acoustics Association conference Forum Acusticum Budapest 2005, pp 1607-1612. Budapest, Hungary: OPAKFI.

Kopco N, Shinn-Cunningham BG (2003) Spatial unmasking of nearby pure-tone targets in a simulated anechoic environment. Journal of the Acoustical Society of America 114:2856-2870.

Kopco N, Carpenter GA (2004) PointMap: A Real-Time Memory-Based Learning System with On-line and Post-Training Pruning. International Journal of Hybrid Intelligent Systems 1:57-71.

Kopco N, Shinn-Cunningham BG (2008) Influences of modulation and spatial separation on detection of a masked broadband target. Journal of the Acoustical Society of America 124:2236-2250.

Kopco N, Schoolmaster M, Shinn-Cunningham BG (2004) Learning to judge distance of nearby sounds in reverberant and anechoic environments. In: Joint Congress of CFADAGA ‘04. Strassbourg, France.

Kopco N, Best V, Shinn-Cunningham BG (2007a) Sound localization with a preceding distractor. J Acoust Soc Am 121:420-432.

Kopco N, Lin I-F, Groh JM, Shinn-Cunningham BG (2007b) Visually-induced auditory spatial adaptation in monkeys and humans. In: Society for Neuroscience Abstract No. 662.4. San Diego, CA.

Lane C, Kopco N, Delgutte B, Shinn-Cunningham B, Colburn H (2004) A cat's cocktail party: Psychophysical, neurophysiological, and computational studies of spatial

release from masking. In: Auditory Signal Processing: Physiology, Psychoacoustics, and Models (Pressnitzer D, Cheveigne Ad, McAdams S, Collet L, eds), pp 405-413. Dourdan, France: Springer Verlag.

Lin I-F, Kopco N, Groh JM, Shinn-Cunningham BG (2007) Characteristics of visually-induced auditory spatial adaptation. Journal of the Acoustical Society of America 121:3095.

Lutfi RA (1990) How much masking is informational masking? Journal of the Acoustical Society of America 88:2607-2610.

Marsalek P (2001) Neural code for sound localization at low frequencies. Neurocomputing 38-40:1443-1452.

Marsalek P, Kofranek J (2005) Sound localization at high frequencies and across the frequency range. Neurocomputing 58-60:999-1006.

Mellinger DK, Mont-Reynaud BM (1996) Scene Analysis. In: Auditory Computation (Hawkins HL, McMullen TA, Popper AN, Fay RR, eds), pp 271-331. New York: Springer Verlag.

Middlebrooks JC, Green DM (1991) Sound localization by human listeners. Annual Review of Psychology 42:135-159.

Moore BCJ (1997) An Introduction to the Psychology of Hearing (4e). San Diego, CA: Academic Press.

Oxenham AJ, Fligor BJ, Mason CR, Kidd G, Jr. (2003) Informational masking and musical training. J Acoust Soc Am 114:1543-1549.

Plomp R, Steeneken HJM, Houtgast T (1980) Predicting speech intelligibility in rooms from the modulation transfer function II. Mirror image computer model applied to rectangular rooms. Acustica 46:73-81.

Saberi K, Dostal L, Sadralodabai T, Bull V, Perrott DR (1991) Free-field release from masking. Journal of the Acoustical Society of America 90:1355-1370.

Sach A, Hill N, Bailey P (2000) Auditory spatial attention using interaural time differences. Journal of Experimental Psychology: Human Perception and Performance 26:717-729.

Santon F (1987) Détection d'un son pur dans un bruit masquant suivant l'angle d'incidence du bruit. Relation avec le seuil de réception de la parole (Detection of a pure sound

in the presence of masking noise, and its dependence on the angle of incidence of noise. Relation with the speech reception threshold). Acustica 63:222-228.

Schickler J, Kopco N, Shinn-Cunningham BG, Litovsky RL (2000) Spatial unmasking of nearby speech sources in a simulated anechoic environment. Journal of the Acoustical Society of America 107:2849.

Schoolmaster M, Kopco N, Shinn-Cunningham BG (2003) Effects of reverberation and experience on distance perception in simulated environments. Journal of the Acoustical Society of America 113:2285.

Shinn-Cunningham BG (2008) Object-based auditory and visual attention. Trends in Cognitive Sciences:in press.

Shinn-Cunningham BG, Santarelli S, Kopco N (2000) Tori of confusion: binaural localization cues for sources within reach of a listener. J Acoust Soc Am 107:1627-1636.

Shinn-Cunningham BG, Desloge JG, Kopco N (2001) Empirical and modeled acoustic transfer functions in a simple room: Effects of distance and direction. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp 183-186. New Pfalz, New York.

Shinn-Cunningham BG, Kopco N, Martin TJ (2005) Localizing nearby sound sources in a classroom: Binaural room impulse resonses. Journal of the Acoustical Society of America 117:3100-3115.

Spence CJ, Driver J (1994) Covert spatial orienting in audition: Exogenous and endogenous mechanisms. Journal of Experimental Psychology: Human Perception and Performance 20:555-574.

Stern RM, Colburn HS (1978) Theory of binaural interaction based on auditory-nerve data. IV. A model for subjective lateral position. Journal of the Acoustical Society of America 64:127-140.

Strutt JW (1907) On Our Perception of Sound Direction. Philosophical Magazine 13:214-232.

Treisman A (1971) Shifting attention between the ears. Quarterly Journal of Experimental Psychology 23:157-167.

Treisman AM, Davies A (1973) Divied attention to ear and eye. In: Attention and Performance (Kornblum E, ed), pp 101-117. New York, NY: Academic Press.

van de Par S, Kohlrausch A (1999) Dependence of binaural masking level differences on center frequency, masker bandwidth, and interaural parameters. J Acoust Soc Am 106:1940-1947.

Viemeister NF (1979) Temporal modulation transfer functions based upon modulation thresholds. Journal of the Acoustical Society of America 66:1364-1380.

Watson CS, Kelly WJ, Wroton HW (1976) Factors in the discrimination of tonal patterns II: Selective attention and learning under various levels of stimulus uncertainity. Journal of the Acoustical Society of America 60:1176-1186.

Wrigley SN, Brown GJ (2001) A Neural Oscillator Model of Auditory Attention. In: Proceedings of the International Conference on Artificial Neural Networks. Lecture Notes In Computer Science; Vol. 2130, pp 1163-1170.

Yantis S (2005) How visual salience wins the battle for awareness. Nature Neuroscience 8:975-977.

Zurek PM (1993) Binaural advantages and directional effects in speech intelligibility. In: Acoustical Factors Affecting Hearing Aid Performance (Studebaker G, Hochberg I, eds). Boston, MA: College-Hill Press.

Spatial unmasking of nearby pure-tone targets in a simulated anechoic environment

Norbert Kopčo and Barbara G. Shinn-Cunningham[a)]
Hearing Research Center, Boston University, Boston, Massachusetts 02215

(Received 31 December 2002; revised 12 August 2003; accepted 15 August 2003)

Detection thresholds were measured for different spatial configurations of 500- and 1000-Hz pure-tone targets and broadband maskers. Sources were simulated using individually measured head-related transfer functions (HRTFs) for source positions varying in both azimuth and distance. For the spatial configurations tested, thresholds ranged over 50 dB, primarily as a result of large changes in the target-to-masker ratio (TMR) with changes in target and masker locations. Intersubject differences in both HRTFs and in binaural sensitivity were large; however, the overall pattern of results was similar across subjects. As expected, detection thresholds were generally smaller when the target and masker were separated in azimuth than when they were at the same location. However, in some cases, azimuthal separation of target and masker yielded little change or even a small increase in detection threshold. Significant intersubject differences occurred as a result both of differences in monaural and binaural acoustic cues in the individualized HRTFs and of different binaural contributions to performance. Model predictions captured general trends in the pattern of spatial unmasking. However, subject-specific model predictions did not account for the observed individual differences in performance, even after taking into account individual differences in HRTF measurements and overall binaural sensitivity. These results suggest that individuals differ not only in their overall sensitivity to binaural cues, but also in how their binaural sensitivity varies with the spatial position of (and interaural differences in) the masker. [DOI: 10.1121/1.1616577]

PACS numbers: 43.66.Pn, 43.66.Ba, 43.66.Qp [LRB] Pages: 2856–2870

I. INTRODUCTION AND BACKGROUND

When listening for a target sound in the presence of a masking sound, a listener's ability to detect the target is influenced by the locations of both target and masker. When target and masker are at the same distance, it is generally easier to detect or recognize the target when it is spatially separated from the masker compared to when the target and masker are at the same position. This "spatial unmasking" effect has been studied for many types of stimuli, including speech (e.g., see Freyman *et al.*, 1999; Shinn-Cunningham *et al.*, 2001), click-trains (e.g., see Saberi *et al.*, 1991; Good *et al.*, 1997), and tone complexes (e.g., see Kidd *et al.*, 1998).

For broadband noise maskers, spatial unmasking arises primarily from acoustic "better-ear" effects (moving a sound source in space changes the levels of the signal reaching the ears of the listener) and "binaural" effects. "Better-ear" effects lead to unmasking because the target-to-masker ratio (TMR) generally increases at one ear when target and masker are in different directions compared to when they are in the same direction. Binaural unmasking can occur when the interaural time and intensity differences in the target and masker differ.

There have been many studies of how binaural differences affect tone detectability in noise [see Durlach and Colburn (1978) for a review of this classic literature]. However, most of these studies were performed under headphones using interaural differences that do not occur naturally. There are only a few studies that have measured how tone detection is affected by the spatial locations of target and masker (examples include Ebata *et al.*, 1968; Gatehouse, 1987; Santon, 1987; Doll and Hanna, 1995). Moreover, results of these studies are inconsistent, finding spatial unmasking ranging from as little as 7 or 8 dB [Santon (1987) and Doll and Hanna (1995), respectively] to as much as 24 dB (Gatehouse, 1987). These apparent discrepancies may be caused by differences in the spatial configurations tested. However, none of these studies analyzed how the TMR at the ears changed with spatial configuration and did not factor out how better-ear (versus binaural) factors may have contributed to the observed spatial unmasking.

Previous studies of spatial unmasking for pure-tone targets considered sources relatively far from the listener and looked only at unmasking resulting from changes in source direction, ignoring any effects of source distance. For sources more than about a meter from the listener, the only significant effect of changing source distance is a change in signal level that is equal at the two ears. However, changes in source distance for sources within reach of the listener produce changes in signal level that differ at the two ears, resulting in exceptionally large interaural level differences (ILDs; see Brungart and Rabinowitz, 1999; Shinn-Cunningham *et al.*, 2000), even at low frequencies for which ILDs are essentially zero for relatively distant sources. In addition, for near sources, relatively small positional changes

a)Author to whom correspondence should be addressed: Department of Cognitive and Neural Systems, Boston University, 677 Beacon St., Room 311, Boston, MA 02215. Electronic mail: shinn@cns.bu.edu

can lead to large changes in the energy of the target and masker reaching the two ears. A few previous studies hint that, in some conditions, binaural performance can be worse than monaural performance using the better ear, particularly when there are large ILDs in the stimuli (e.g., see Bronkhorst and Plomp, 1988; Shinn-Cunningham *et al.*, 2001). Given that large ILDs can arise when sources are within reach of the listener, studies of binaural unmasking for nearby sound sources may shed light on these reports.

The current study examined spatial unmasking of pure tone sources within reach of a listener in a simulated anechoic environment. Individually measured head-related transfer functions (HRTFs) were used to simulate sources. This approach allowed realistic spatial acoustic cues to be presented to the subjects while still allowing detailed analyses of the stimuli reaching the subjects during the experiment. The main goals of the study were to (1) measure how target threshold depends on target and masker azimuth and distance for nearby sources, (2) characterize better-ear effects by analyzing how the TMR varies with the spatial configurations tested, (3) evaluate the binaural contribution to spatial unmasking, particularly for spatial configurations in which large ILDs arise, and (4) investigate the degree to which results can be accounted for by a model of binaural interaction.

II. SPATIAL UNMASKING OF NEARBY PURE TONE TARGETS

A. Methods

1. Subjects

Four graduate students with prior experience in psychoacoustic experiments (including author NK) participated in the study. One subject was female and three were male. Subject ages ranged from 25 to 28 years. All subjects had normal hearing as confirmed by an audiometric screening.

2. HRTF measurement

Individualized HRTF measurements were made with subjects seated in the center of a quiet classroom (rough dimensions of 5×9×3.5 m; broadband T_{60} of approximately 700 ms). Subjects were seated with their heads in a headrest so that their ears were approximately 1.5-m above the floor. Measurements were taken for sources in the right front horizontal plane (at ear height) for all six combinations of azimuths (0°, 45°, 90°) and distances (0.15 m, 1 m) relative to the center of the head (defined as the intersection of the interaural axis and the median plane) as shown in Fig. 1.

The Maximum-Length-Sequence (MLS) technique (e.g., see Vanderkooy, 1994) was used to measure HRTFs. Two identical 32 767-long maximum length sequences were concatenated and presented through a small loudspeaker using a 44.1-kHz sampling rate (details regarding the equipment are described below). The response to the second sequence was recorded.[1] This measurement was repeated ten times and the raw measurements averaged in the time domain. This average response was then used to estimate a 743-ms-long head-related impulse response.

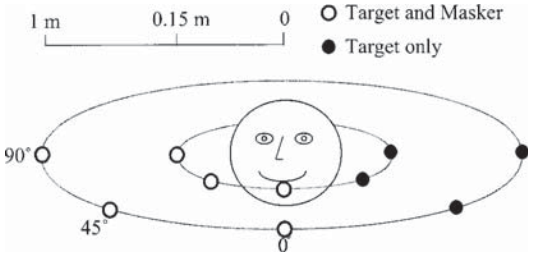

FIG. 1. Spatial positions used in the study. HRTFs were measured at the positions denoted by open symbols. Target detection thresholds were measured for all spatial combination of six masker positions (open symbols) and ten target positions (filled and open symbols; targets simulated at the filled symbols used the corresponding HRTFs from the contralateral hemifield with left- and right-ear signals reversed).

HRTFs were measured using a Tucker-Davis Technologies (TDT) signal processing system under computer control. For each measurement, the concatenated MLS sequence was read from a PC hard-drive and sent to a TDT D/A converter (TDT PD1), which drove a Crown amplifier connected to a BOSE mini-cube loudspeaker. At the start of the measurement session, the subject was positioned so that the center of his/her head was at a location marked on the floor of the room. The subject's head position was read from a Polhemus FastTrak electromagnetic tracker worn on the head to ensure that the center of the head was within 1-cm of the correct location in the room, marked on the floor. The experimenter used other angular and distance markings on the floor to hand-position the loudspeaker to the appropriate azimuth and distance prior to each measurement. Miniature microphones (Knowles FG-3329c) mounted in earplugs and inserted into the entrance of the subjects' ear canals (to produce blocked-meatus HRTF recordings) measured the raw acoustic responses to the MLS sequence. Microphone outputs drove a custom-built microphone amplifier that was connected to a TDT A/D converter (TDT PD1). These raw results were stored in digital form on the computer hard-drive for off-line processing to produce the estimated HRTFs.

No correction for the measurement system transfer function was performed, but the amplitude spectrum of the transfer-function of this measurement system was examined and found to vary by less than 2 dB and to cause no significant interaural distortion for frequencies between 400 and 1500 Hz (the frequency region important for the current study). The useful dynamic range of the measurements (taking into account the ambient acoustic and electrical noise) was at least 50 dB for all frequencies greater than 300 Hz.

HRTFs measured as described above include room echoes and reverberation. To eliminate room effects, time-domain impulse responses were multiplied by a 6-ms-long cos-squared time window (rise/fall time of 1 ms) to exclude all of the reverberant energy while retaining all of the direct-sound energy. The resulting ''pseudo-anechoic'' HRTFs were used to simulate sources (and in all subsequent analyses).

HRTFs were measured only for sources in the right hemifield. To simulate sources in the left hemifield, HRTFs from the corresponding right-hemifield position were used, exchanging the left and right channels (i.e., left/right symmetry was assumed; given that only pure tone targets were

simulated in the left hemifield, this approximation should introduce no significant perceptual artifacts in the simulated stimuli).

The measured HRTFs reflect the radiation characteristics of the loudspeaker used, which is not a uniformly radiating point source. For sources relatively far from the head, any differences in the measurement caused by the directivity of the source should be minor. For sources 15-cm from the center of the head, the effect of the source directivity may be significant. Therefore, the current study focuses on how distance influenced the signals reaching the ears for the particular source used (the Bose loudspeaker in question). The issue of how well the current results may generalize to other nearby sources is considered further in Sec. III, where empirical HRTF measurements are compared with theoretical predictions from a spherical head model that assumes a perfect point source.

In a similar vein, HRTFs measured for sources close to the head are much more sensitive to small displacements in the source (*re*: the intended source location) than more distant sources. However, given that all acoustic analyses and predictions of performance were made using the same measured HRTFs used to simulate the headphone-presented stimuli, any conclusions regarding which acoustic factors influence performance are justified, even if other measurement techniques might yield slightly different estimates of near-source HRTFs for the positions reported here.

3. Stimulus generation

Target stimuli consisted of 165-ms-long pure tones of 500 or 1000 Hz gated on and off by 30-ms cos-squared ramps. The 500-Hz target frequency was chosen so that results could be compared with previous studies of binaural masking level differences (BMLDs) and spatial unmasking of tones, most of which include a 500-Hz target condition. The 1000-Hz target was included in order to examine what happens for a higher target frequency where target and masker ITDs are still likely to have a large impact on detection but ILDs are larger than at 500 Hz. The target was temporally centered within a broadband, 250-ms-long masker. On each trial, the masker token was randomly chosen from a set of 100 pregenerated samples of broadband noise that were digitally low-pass filtered with a 5000 Hz cutoff frequency (ninth-order Butterworth filter, as implemented in the signal-processing toolbox in Matlab, the Mathworks, Natick, MA).

In most cases, target and masker were simulated as arising from different locations in anechoic space by convolving the stimuli with appropriate individualized head-related impulse responses (time-domain representation of the HRTFs). The simulated spatial configurations included all combinations of target at azimuths (−90°, −45°, 0°, 45°, 90°) and distances (0.15 m, 1 m) and masker at azimuths (0°, 45°, 90°) and distances (0.15 m, 1 m). A total of 60 spatial configurations was tested (10 target locations × 6 masker locations; see Fig. 1). In a subset of trials, traditional BMLDs were measured using the same stimuli without HRTF processing.

For nearby sources, keeping the masker presentation level constant would result in the received level (at the subject's ears) varying widely with masker position. In order to keep the received level of masker relatively constant, the levels of the HRTF-processed masker stimuli were normalized to keep constant the rms energy falling within the equivalent rectangular band (ERB; Moore, 1997) centered on the target frequency at the ear receiving the more intense masker signal (the right ear for all of the tested configurations). In other words, the virtual stimuli actually simulated a masker whose distal energy level was adjusted up or down (depending on the masker spatial location) until the proximal stimulus level was constant at the more intense ear. In our analysis, the amounts by which the distal masker was adjusted were added back to the raw thresholds to predict the amount of spatial unmasking that would have occurred if the distal masker level had been constant.[2]

For the 500-Hz center frequency, the rms levels were adjusted using a 100-Hz-wide ERB. For the 1000-Hz target, the ERB width was set to 136 Hz. The masker signals were preprocessed in Matlab so that the right- (more-intense-) ear rms masker level in the ERB would be 64 dB SPL when played via headphones. BMLDs were measured with the low-pass-filtered noise spectral level fixed at 64 dB SPL.

Stimulus files, generated at a sampling rate of 44.1 kHz, were stored on the hard disk of the control computer (IBM PC compatible). On each trial, appropriate target and masker signals were presented through TDT hardware. Left- and right-ear target and masker signals were processed through four separate D/A converters (TDT PD1). Target signals were scaled to the appropriate presentation level by a programmable attenuator (TDT PA4), summed with the fixed-level masker signals (TDT SM3), and amplified through a headphone buffer (TDT HB6). The resulting binaural stimuli were presented via Etymotic Research ER-1 insert earphones. No filtering was done to compensate for the transfer characteristics of the playback system. A handheld RS 232 terminal (QTERM) was used to gather subject responses and provide feedback.

4. Experimental procedure

Behavioral experiments were performed in a single-walled sound-treated booth.

Each trial consisted of three intervals, each of which contained a noise burst. Either the second or third interval (randomly chosen, with equal probability, on each trial) also contained the tone-burst target. Subjects performed a two-alternative, forced-choice task in which they were asked to identify which interval, the second or third, contained the target tone. Correct-answer feedback was provided at the end of each trial.

A three-down–one-up adaptive procedure was used to estimate detection thresholds (Levitt, 1971), defined as the 79.4% correct point on the psychometric function. Each run started with the target at a clearly detectable level and continued until 11 "reversals" occurred. The target level was changed by 4 dB on the first reversal, 2 dB on the second reversal, and 1 dB on all subsequent reversals. For each adaptive run, detection threshold was estimated by taking the average target presentation level over the last six reversals.

TABLE I. Binaural masking level differences for individual subjects. Note that subjects S1 and S3 performed detection experiments for both 500- and 1000-Hz targets; S2 and S4 only performed the experiments for one target frequency (500 and 1000 Hz, respectively). Symbols give the convention used in the figures when plotting individual subject results.

Target frequency	Individual subject results				Across-subject average
	S1 ○	S2 ▽	S3 □	S4 △	
500 Hz	15.6	11.0	14.5	NA	13.7
1000 Hz	13.1	NA	7.5	8.7	9.8

At least three separate runs were performed for each subject in each condition. Final threshold estimates were computed by taking the average threshold across the repeated adaptive threshold estimates. Additional adaptive runs were performed as needed for every subject and condition to ensure that the standard error in this final threshold estimate was less than or equal to 1 dB for each condition and spatial configuration tested.

The study was divided into two parts, one measuring thresholds for the 500-Hz target and one for the 1000-Hz target. Three subjects performed each part (two of the four subjects performed both). For each target, subjects performed multiple sessions consisting of ten runs. Subjects were allowed to take short breaks between runs within one session, with a minimum 4-h break required between sessions. Each subject performed one initial practice session consisting of four practice runs and six runs measuring detection thresholds for NoSo and NoSπ conditions (where NoSo represents a sinusoidal diotic signal, i.e., with zero interaural phase difference, in the presence of a diotic noise; NoSπ represents a sinusoidal signal with interaural phase difference equal to π in the presence of a diotic noise). Subjects then performed 18 additional sessions (180 runs; 3 runs each of every combination for 6 target positions and 10 masker positions). In each of these sessions, a full set of thresholds was determined for one masker position (the order of the ten target positions was randomized within each session). These sessions were grouped into three blocks of six with each block containing a full set of thresholds. The order of masker positions was separately randomized for each block and subject. Any additional runs were performed after completion of the initial 19 sessions. Each subject performed approximately 20 h of testing per target frequency.

B. Results

1. Binaural masking level difference

Table I shows the BMLD (see Durlach and Colburn, 1978), defined as the difference in target detection threshold in the NoSo and NoSπ conditions. Results are consistent with those from previous, similar experiments. BMLDs are larger for the 500-Hz target (where BMLDs ranged from 11 to 16 dB) than the 1000-Hz target (where BMLDs ranged from 7 to 14 dB).

2. Spatial unmasking

The amount of "spatial unmasking" is defined as the change in the energy a target emits at threshold for a particular target location compared to when the target is at the same position as the masker. In order to estimate the target detection threshold when the emitted level of the masker is held constant, the amount by which the masker was normalized (to equate the masker level at the more intense ear) was first added back to the raw target detection thresholds. To estimate spatial unmasking (i.e., the amount by which detection thresholds improve with spatial separation of target and masker), the average of all thresholds when target and masker were at the same location was computed and this value was subtracted from all the renormalized thresholds.

Figures 2 and 3 plot the amount of spatial unmasking for 500- and 1000-Hz targets, respectively. Each panel shows the amount of spatial unmasking (improvement in target threshold relative to when target and masker are at the same location) for one masker location (shown graphically in the inset legend in each panel). The abscissa shows the target azimuth. Thick lines and filled symbols show results for the near target; thin lines and open symbols show results for the far target. Symbols show individual subject results and solid lines give the across-subject mean. Dashed lines represent the estimates of the better-ear contribution to spatial unmasking (averaged across subjects), discussed in detail in Sec. IV.

For the spatial configurations tested, the amount of spatial unmasking spans a range of over 50 dB [e.g., compare the thresholds for a 500-Hz target at (0°, 1 m), the center of the thin line in Fig. 2(d), to the thresholds for the 500-Hz target at (90°, 15 cm), the rightmost point of the thick line in Fig. 2(a)]. While subjects generally show similar patterns of results, intersubject differences are large. For instance, in Fig. 2(a) when the masker is at (0°, 1 m) and the 500-Hz target is at 15-cm, subject S1 (filled circles) consistently shows as much as 10 dB more unmasking than the other subjects (other filled symbols). However, this same subject consistently shows the least unmasking in other cases [e.g., in Fig. 2(f) when the masker is at (90°, 15 cm) and the target is at 1-m; compare open circles to the other open symbols].

Despite the large intersubject differences, overall trends are similar across subjects and for both 500- and 1000-Hz targets, and are summarized below.

To a first-order approximation, changing either target or source distance influences spatial unmasking in a straightforward way predicted by a simple change in the stimulus levels at the ears. For instance, looking within any single panel in Fig. 2 or 3 shows that positioning the target near the subject (thick lines) improves target detectability compared to when the target is far from the subject (thin lines; i.e., within any single panel thick lines are grossly similar to thin line results shifted upward by 10–20 dB). Similarly, comparison of the upper panels (a, b, and c) to the lower panels (d, e, and f) shows that positioning the masker near the subject (lower panels) degrades target detectability compared to when the masker is farther from the subject (upper panels; i.e., results in the upper panels are grossly similar to results in the lower panels shifted upward by 10–15 dB). However, closer inspection shows that the detailed pattern of spatial unmasking varies in a more complex way with both target and masker distance than a simple shift in threshold.

Spatial unmasking resulting from a fixed angular separation of target and masker is larger for nearby targets than

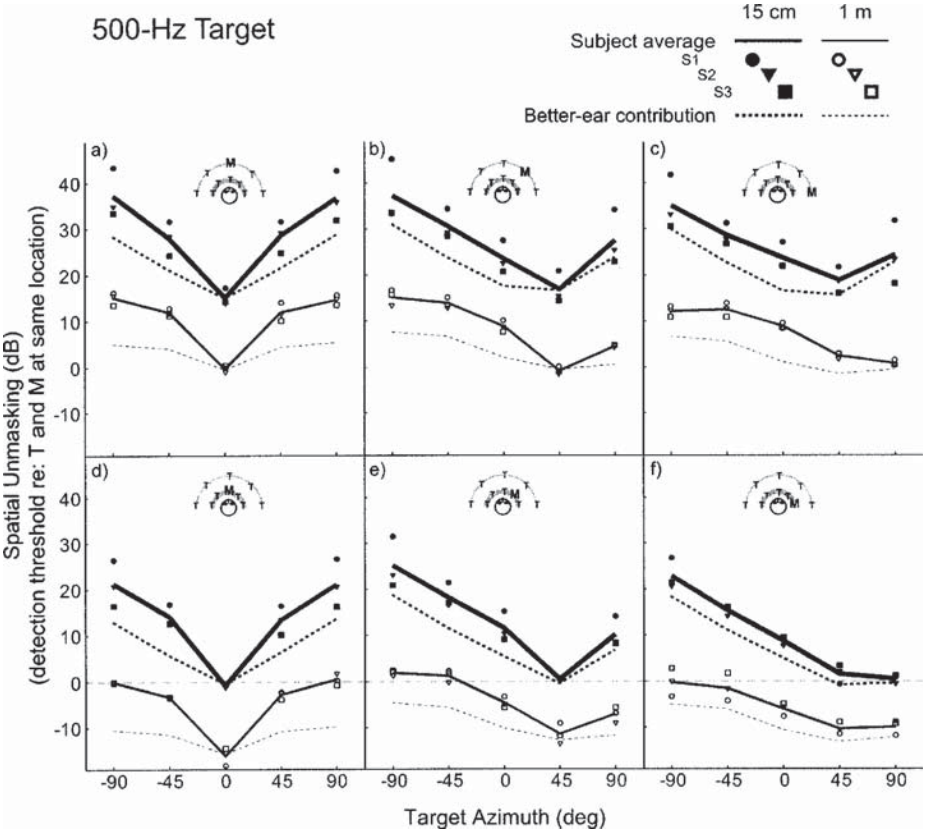

FIG. 2. Spatial unmasking for the 500-Hz target. Each panel plots spatial unmasking (the difference between target detection threshold when target and masker are at the same spatial location and when target and masker are in the spatial configuration denoted in the plot) as a function of target azimuth for a fixed masker location. Across-subject averages are plotted for target distances of 15-cm (thick solid lines) and 1-m (thin solid lines). Individual subject results are plotted as symbols. Dashed lines show the estimated better-ear contribution to spatial unmasking. The spatial configurations of target and masker represented in each panel are denoted in the panel legend.

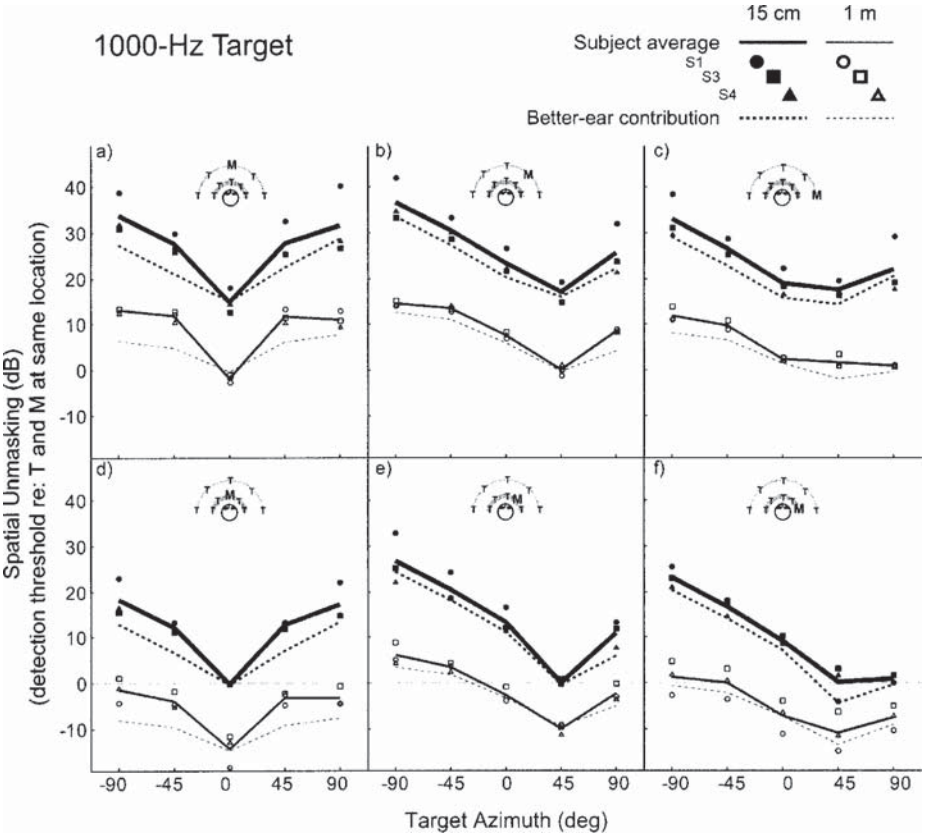

FIG. 3. Spatial unmasking for the 1000-Hz target. See caption for Fig. 2.

for distant targets. For example, in Fig. 3e, the difference between thresholds for the −90° and 45° targets is more than 25 dB for nearby targets (thick line) but less than 20 dB for distant targets (thin line).

Similarly, spatial unmasking resulting from a fixed angular separation of target and masker is larger for nearby maskers than for distant maskers. For example, as discussed above, for a 1000-Hz target when the masker is at (45°, 15 cm) [Fig. 3(e)], spatial unmasking for a 15-cm target (thick line) decreases by more than 25 dB when the target azimuth changes from −90° to +45°. However, when the masker is at (45°, 1 m) [Fig. 3(b)], this same angular displacement of the 15-cm target (thick line) produces a change in spatial unmasking of roughly 20 dB (compare the leftmost point and the point producing the least spatial unmasking, where the target is at 45°).

Angular separation of target and masker can actually make performance worse when target distance differs from masker distance. Usually, separating target and masker in azimuth improves target detectability compared to when the target and masker are in the same direction, but not in every case. When the masker is at 0° (panels a and d in both Figs. 2 and 3) the least amount of spatial unmasking occurs (thresholds are highest) when the target is at 0° (the same direction as the masker); when the masker is at 45° (panels b and e in Figs. 2 and 3) the least unmasking arises when the target is in the 45° masker direction. However, when the masker is at 90° (panels c and f in Figs. 2 and 3), angular separation of target and masker does not always increase the amount of unmasking. Specifically, for a masker at (90°, 1 m) [Figs. 2(c) and 3(c)] there is less spatial unmasking when the 15-cm target (thick line) is at 45° than when it is at 90°. Similarly, for a masker at (90°, 15 cm) [Figs. 2(f) and 3(f)] the amount of spatial unmasking for a 1-m target (thin line) is either equal [500-Hz target; Fig. 2(f)] or greater [1000-Hz target; Fig. 3(f)] when the target is at 90° compared to 45°.

Finally, independent of target or masker distance, the same angular separation of target and masker tends to produce less spatial unmasking as the masker laterality increases. For example, in Fig. 2(d) when the masker is at (0°, 15 cm) and the 500-Hz target is at a distance of 15-cm (thick line), a 90° angular separation of target and masker yields nearly 20 dB of unmasking. However, in Fig. 2(f), when the masker is at (90°, 15 cm) and the target is at 15-cm (thick line), the same angular separation of target and masker produces only 10 dB of unmasking.

C. Discussion

Intersubject differences in spatial unmasking may be partially explained by intersubject differences in the size of the BMLD. For instance, subject S1 has the largest BMLDs and exhibits the most spatial unmasking. However, intersubject differences in spatial unmasking could also be caused by differences in the acoustic parameters in the individually measured HRTFs. Analysis of acoustic differences in the measurements and the binaural contribution to spatial unmasking, which are considered further in Sec. IV, suggest that intersubject differences in spatial unmasking are affected both by subject-specific differences in acoustic cues and in different sensitivities to binaural cues.

Many of the current results follow easily predicted patterns. Moving the target closer to the subject improves detection performance (as expected on the basis of an increase in the level of the target reaching the listener); conversely, moving the masker closer degrades detection performance (as expected when the level of the masker at the ears increases). Separating target and masker in angle improves detection performance for most spatial configurations. However, there are other effects that are less intuitive. Unmasking varies more with target azimuth for a 15-cm masker than for a 1-m masker and for a 15-cm target than for a 1-m target. The masker laterality influences the effectiveness of a given angular separation of target and masker, decreasing with masker laterality. Finally, when target and masker are at different distances and the masker is at 90°, the amount of unmasking can actually decrease when the target is at 45° compared to when the target is in the same direction as the masker (this is essentially a case where there is "spatial masking," i.e., where performance is actually worse when the sources are spatially separated compared to when they are at the same location).

Apparent discrepancies in the amount of spatial unmasking observed in previous studies are actually consistent with the current results. For example, the current study found more spatial unmasking for 1-m sources when the masker is at 0° compared to when the masker is at 90°. Thus, the relatively large amount of spatial unmasking observed by Gatehouse (1987) compared to that found by Santon (1987) and Doll and Hanna (1995) may be caused by the fact that Gatehouse fixed the masker in front of the listener and varied target azimuth, whereas Santon and Doll and Hanna fixed the target in front of the listener and varied masker azimuth.

III. HRTF MEASUREMENTS

The acoustic factors that influence spatial unmasking can be characterized by analysis of the HRTFs used in the simulations. Three acoustic characteristics of the HRTFs influence the performance in a spatial unmasking task: the magnitude spectra of, the interaural level differences (ILDs) in, and the interaural time differences (ITDs) in the signals reaching the two ears. The magnitude spectra of the HRTFs determine the intensity of the sound at the ears and thus the amount of spatial unmasking resulting from better-ear effects. ITDs and ILDs determine the amount of binaural unmasking. In this section, these parameters are analyzed for the individually measured HRTFs.

Individual HRTFs for the four human subjects are compared both to values measured for a KEMAR acoustic manikin (using the same measurement techniques used for the individual subjects) and those predicted from a spherical model of the head assuming a perfect point source. While the literature contains descriptions of both KEMAR (Brungart and Rabinowitz, 1999) and spherical-head model (Duda and Martens, 1998; Shinn-Cunningham *et al.*, 2000) HRTFs for sources near the listener, the current analysis compares these "generic" models to human measurements to determine

whether the models capture the acoustic effects that are important for predicting the amount of spatial unmasking as a function of nearby target and masker locations. As noted in Sec. II, the current measurements do not try to compensate for the radiation characteristics of the loudspeaker used; as such, any consistent discrepancies between predictions from a spherical-head model and measured results (from KEMAR and the human subjects) may reflect influences of the radiation characteristics of the loudspeaker used (which is not a point source) or other differences between the assumptions of the spherical-head model and properties of the physical sources and heads measured.

A. Methods

KEMAR HRTFs were measured using a procedure identical to that used for the human listeners (see description in Sec. II). HRTF predictions for a spherical head model (Brungart and Rabinowitz, 1999; Shinn-Cunningham *et al.*, 2000) were computed using a head with radius of 9-cm and diametrically opposed ears. These results are compared to the HRTFs measured for the four subjects who participated in the spatial unmasking experiment.

For all of the HRTFs, the magnitude spectra, ILD, and ITD were determined for the equivalent rectangular band (ERB) centered at a given frequency. Magnitude spectra were calculated as the rms energy in the HRTF falling within each ERB filter (100-Hz width centered at 500 Hz and 136-Hz width centered at 1000 Hz). ILDs were computed as the difference in the magnitude spectra for the left and right ears. ITD was first estimated as a function of frequency by taking the difference between the right- and left-ear HRTF phase angles at each frequency f and dividing by $2\pi f$. The ITD in each ERB filter was then estimated as the average of the ITD values for the frequencies falling within each ERB filter.

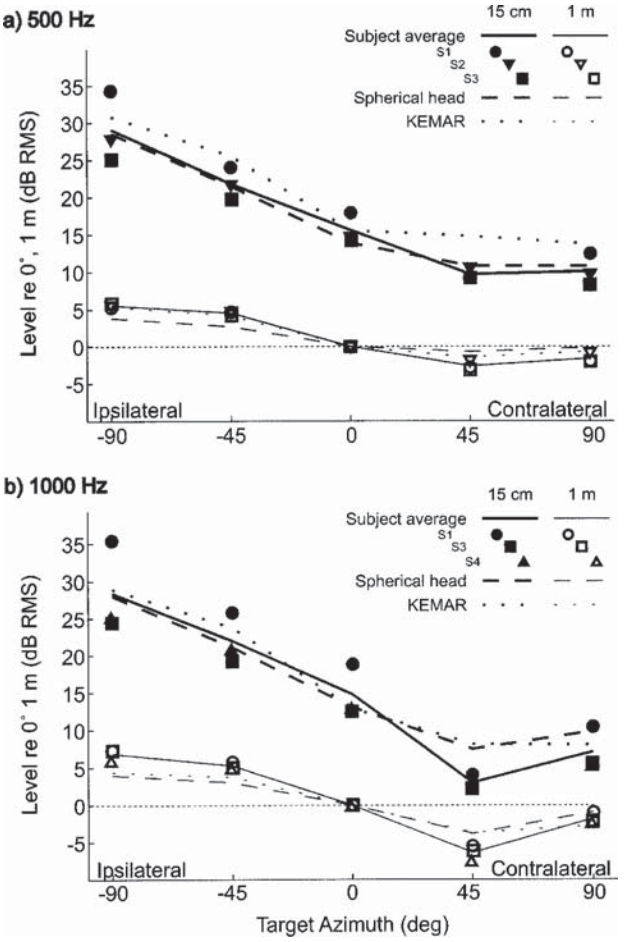

FIG. 4. Left-ear HRTF spectrum levels in ERB filters, relative to the left-ear HRTF for a source at (0°, 1 m). Results are shown for individual listeners, KEMAR, and the spherical head model as a function of source position. (a) 500 Hz. (b) 1000 Hz.

B. Results

1. Intensity effects

Figure 4 shows the magnitude of the ERB-filtered HRTFs at 500 [Fig. 4(a)] and 1000 Hz [Fig. 4(b)] for the left ear relative to a source at (0°, 1 m). (Recall that HRTFs were measured only for sources to the right of the listener and that this analysis assumes left-right symmetry.) Results are shown as a function of the target azimuth for individual human subjects (symbols), the across-human-subject average (solid line), KEMAR (dotted line), and a spherical head model (dashed line). Distant sources are represented by open symbols and thin lines; near sources are shown by filled symbols and thick lines.

Not surprisingly, for both frequencies the spectral gain is larger for near sources (thick lines) than far sources (thin lines). However, in addition to an overall shift in level, the dependence of the HRTF level on source azimuth differs for the two distances. Specifically, for the 15-cm distance (thick lines), the gain to the ipsilateral ear (i.e., the gain for sources at negative azimuths) grows rapidly with source eccentricity compared to the 1-m distance, while the gain to the contralateral ear (positive azimuths) changes similarly with source angle for both distances (compare thick and thin lines).

Overall, intersubject differences are modest for the more distant source (consider the open symbols in each panel). However, there are larger intersubject differences for the 15-cm source positions (filled symbols). For instance, at both frequencies [Figs. 4(a) and (b)], the 15-cm HRTF gain for subject S1 (filled circles) is generally 5–10 dB larger than for the other subjects, except at 45° where all HRTFs are similar.

For a 15-cm source at both 500 Hz [Fig. 4(a)] and 1000 Hz [Fig. 4(b)], KEMAR (thick dotted lines) and spherical-head gains (thick dashed lines) generally fall within the range of values observed for the four human subjects (filled symbols) measured in this study. However, in Fig. 4(b) for a 1-m source, KEMAR measurements (thin dotted lines) and model predictions (thin dashed lines) slightly underestimate the 1000-Hz gain to the ipsilateral ear compared to the individual subject results (lines fall below symbols for azimuths of −45° and −90°). At 500-Hz [Fig. 4(a)], the 1-m KEMAR measurements (thin dotted lines) fall within the range of results obtained from the human subjects (open symbols); however, the spherical head model results (thin dashed lines) fall below the subject measurements (open symbols) for ipsilateral sources (sources at −45° and −90°).

While, intuitively, we expect the level of the signal

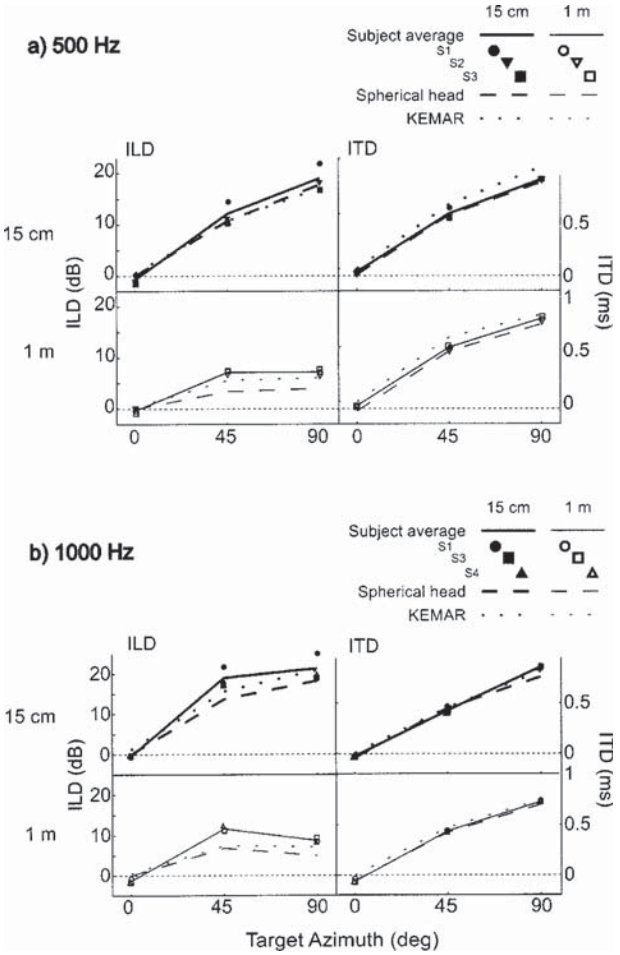

FIG. 5. ILDs and ITDs in HRTFs for individual subjects, KEMAR manikin, and the spherical head model. (a) 500 Hz. (b) 1000 Hz.

reaching the ears to vary monotonically with lateral angle of the source, human HRTF measurements show that this is not strictly true. In particular, the 1000-Hz human measurements [symbols and solid lines in Fig. 4(b)] show that less energy reaches the contralateral ear when a source is at 45° than when it is at 90° for both source distances (thick and thin lines are nonmonotonic with azimuth) Similarly, at 500 Hz [Fig. 4(a)] the gain to the contralateral ear is comparable for 45° and 90° sources rather than decreasing for the 90° source (thick and thin lines). This nonmonotonicity [which may in part be a consequence of the acoustic ''bright spot;'' e.g., see Brungart and Rabinowitz (1999)] is underestimated in both the spherical-head model (dashed lines) and KEMAR (dotted lines) HRTFs, especially at 1000 Hz [compare lines to human subject results for sources at 45°, especially in Fig. 4(b)].

2. Interaural differences

Figure 5 shows the ILDs and ITDs in the measured HRTFs at 500 and 1000 Hz [Figs. 5(a) and (b), respectively] for the spatial positions used in the study. As in Fig. 4, results for individual subjects (symbols), the across-human-subject average (full lines), KEMAR (dotted lines), and a spherical head model (dashed lines) are shown as a function of target azimuth. Results for near sources are shown in the top of each subplot with heavy lines and filled symbols. Thin lines and open symbols plot results for far sources (bottom row of each half of the figure). The left column shows ILD results and the right column shows ITD results.

ILDs were calculated directly from the measurements plotted in Fig. 4. As a result, there are large intersubject differences in the ILDs (left panels in Fig. 5) that are directly related to the intersubject differences in the monaural spectral gains. For instance, subject S1 has much larger ILDs at both 500 and 1000 Hz for the 15-cm source [filled circles in the left columns of Figs. 5(a) and (b)] than any of the other subjects (other filled symbols).

As expected, for both frequencies [Figs. 5(a) and (b)] ILDs are much larger for sources at 15-cm (thick lines in top left panels) compared to 1-m (thin lines in the bottom left panels) with ILDs at 500 and 1000 Hz approaching 20 dB for the nearby sources at 90° (rightmost point in the top left panels). The spherical-head (dashed lines) and KEMAR (dotted lines) results tend to underestimate ILDs for lateral sources, although for the 500-Hz, 15-cm sources [Fig. 5(a), top left panel], both spherical-head and KEMAR results are within the range of human observations. Discrepancies between human and model results are most pronounced for a 1000-Hz source at a distance of 1-m [Fig. 5(b), bottom left panel] and are greater for the spherical-head predictions (dashed lines) than KEMAR measurements (dotted lines).

ITDs [the right panels in Figs. 5(a) and (b)] vary primarily with source angle and change only slightly with distance and frequency. For most of the measured locations, both spherical-head and KEMAR results are in close agreement with human measurements.

C. Discussion

Both spherical-head and KEMAR HRTFs provide reasonable approximations to how acoustic parameters in human HRTFs vary with source location. In general, both KEMAR and the spherical head measurements fall within the range spanned by the individual subject measurements. However, both spherical-head predictions and KEMAR measurements slightly overestimate the gain at the contralateral ear when a source is at 45° (especially at 1000 Hz) and tend to modestly underestimate the ILD for sources off midline, particularly at the 1-m distance. These small differences cannot be attributed to loudspeaker characteristics, given that (1) the discrepancies are similar for both KEMAR measurements (using the same loudspeaker) and spherical-head predictions (assuming a perfect point source) and (2) the differences are, if anything, larger for the more distant, 1-m source (where the loudspeaker directivity is less influential) than the nearby source. Thus, we conclude that generic HRTF models capture the important features of the HRTFs measured in human subjects and that the effects of the source transmission characteristics do not strongly influence the signals reaching the ears even for nearby sources, at least for the frequencies considered in the current study.

Intersubject differences in the HRTFs are large, especially for nearby sources. Of the four subjects, one subject showed consistently larger spectral gains and consistently

larger ILDs than the other subjects when the source was at 15-cm. While it is possible that some of the intersubject differences arise from inaccuracies in HRTF measurement (e.g., from hand-positioning the loudspeaker), the fact that one subject has consistently larger gains and ILDs for all nearby source locations suggests that real anatomical differences rather than measurement errors are responsible for the observed effects. It is also interesting to note that the observed intersubject differences are much smaller for the 1-m source, suggesting that intersubject differences in HRTFs are especially important when considering sources very close to the listener.

IV. BETTER-EAR AND BINAURAL CONTRIBUTIONS TO SPATIAL UNMASKING

A. Analysis

For each subject, estimates of the better-ear and binaural contributions to spatial unmasking were derived from the acoustic parameters of the HRTFs and the behavioral thresholds.

The better-ear contribution to spatial unmasking was estimated by calculating the TMR in the ERB filter centered on the target frequency at the better ear for each spatial configuration when target and masker emit the same level (and thus would yield a TMR of zero when at the same location). The resulting TMR predicts the amount by which target thresholds decrease or increase simply because of acoustic effects at the better ear (i.e., if the calculated TMR is +2 dB, it implies that at detection threshold, the intensity of the target at the better ear was 2 dB more for the given spatial configuration than if the target and masker were at the same spatial location; thus, the better-ear contribution for such a configuration is +2 dB). The subject-specific binaural contribution to spatial unmasking was estimated by subtracting the estimated better-ear contribution to spatial unmasking (derived from individually-measured HRTFs) from the individual behavioral estimates of spatial unmasking.

B. Results

1. Better-ear contributions to spatial unmasking

While intersubject differences in the better-ear contribution to spatial unmasking are large, the trends in the across-subject average data capture the important features of the individual data. For brevity, only the across-subject averages are presented in Figs. 2 and 3 for the 500- and 1000-Hz target, respectively, as dashed lines. For all spatial configurations tested, the behaviorally observed amount of spatial unmasking either equals or is larger than the predicted spatial unmasking from better-ear effects (dashed lines fall below or at measured values in all graphs). Thus, even when there are large ILDs in the signals reaching the listener, binaural performance is always better than or equal to predicted performance when listening monaurally with the acoustically better ear.

Better-ear effects account for a large portion of the observed spatial unmasking when target and masker are in the same direction and for the large influence of target and/or masker distance on spatial unmasking. Specifically, the predicted results (dashed lines) are in good agreement with the measured results when the target is at 0° in the left column, at 45° in the middle column, and at 90° in the right column. Generally, angular separation of target and masker increases the better-ear contribution to unmasking (dashed-line predictions generally increase as the target azimuth moves away from the masker azimuth). However, when the masker is at 90° (the right columns in Figs. 2 and 3), better-ear effects either decrease or are roughly the same when the target is at 45° compared to 90° (dashed-line predictions are either constant or decrease as the target azimuth moves from 90° to 45°). Better-ear contributions to unmasking change more with target azimuth when the target is at 15-cm (thick dashed lines) than at 1-m (thin dashed lines), primarily because, for nearby sources, small positional changes cause large changes in the relative distance from source to the better ear.

Finally, differences between mean subject results (solid lines) and predicted better-ear effects (dashed lines) are generally larger for the 500-Hz target (Fig. 2) than the 1000-Hz target (Fig. 3), suggesting that the better-ear contributions to unmasking are relatively more important (i.e., account for a greater portion of the observed amount of spatial unmasking) for the 1000-Hz target than the 500-Hz target. This is true both because the better-ear effects are larger in absolute terms and because the additional spatial unmasking for which better-ear effects cannot account is smaller at 1000 Hz than at 500 Hz.

2. Binaural contributions to spatial unmasking

Figures 6 and 7 show the estimated binaural contribution to spatial unmasking for the 500- and 1000-Hz target, respectively. The binaural contribution was calculated for each individual subject by subtracting the estimated better-ear contribution (the across-subject average of which is shown by dashed lines in Figs. 2 and 3) from the total amount of spatial unmasking (symbols in Figs. 2 and 3). Both Figs. 6 and 7 show results for each subject who performed that condition in a separate subplot. Each subplot is divided into six panels corresponding to the six masker locations (laid out as indicated in the legend). In each panel, symbols plot the mean binaural contribution to spatial unmasking (averaged across the repeated adaptive runs). The error bars show the *range* of thresholds obtained across the repeated adaptive runs for each condition. Results are shown for both the far target (gray) and the near target (black) as a function of target azimuth. Figures 6 and 7 also show model predictions (lines), which are derived and discussed in Sec. V.

Even though intersubject differences are large, there are a number of trends that are consistent across subjects. Not surprisingly, for both target frequencies (Figs. 6 and 7) there is no unmasking beyond the better-ear contribution when target and masker are at the same spatial location (the binaural gain is near zero when the target is at 0° in the left columns, at 45° in the middle columns, and at 90° in the right columns of Figs. 6 and 7). In fact, only the 500-Hz results for subject S1 [Fig. 6(a)] show any binaural unmasking when target and masker are at the same off-median-plane direction but at different distances. For example, looking at the top right panel of Fig. 6(a) [masker at (90°, 1 m)], the binaural gain is posi-

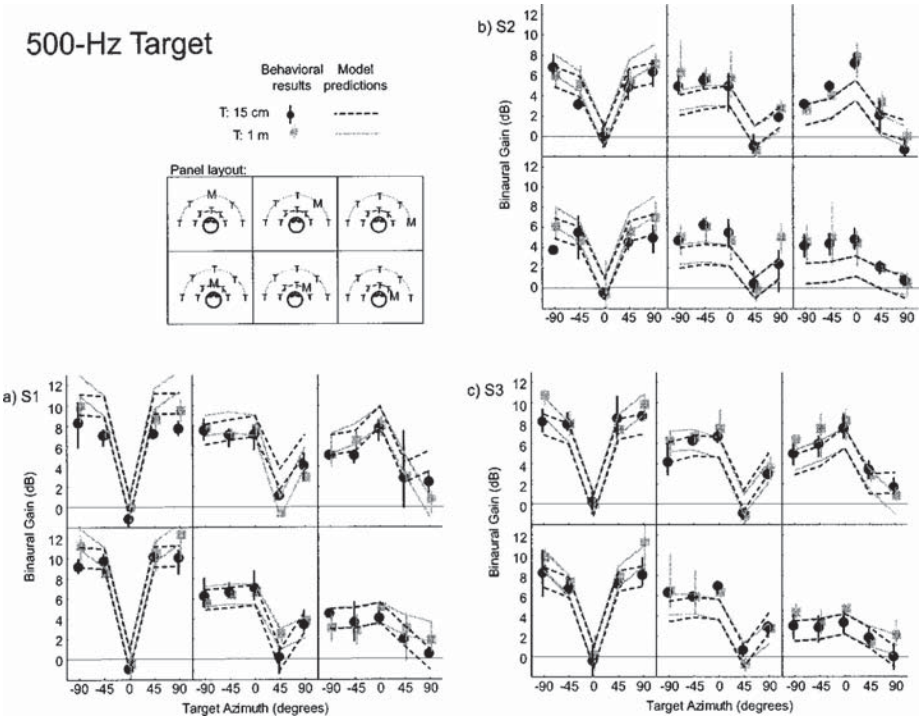

FIG. 6. Estimated binaural contribution to spatial unmasking for the 500-Hz target. Each panel plots the amount of binaural unmasking for one masker position for both the 15-cm and 1-m target. Symbols show estimates for individual subjects with error bars showing the range of results across multiple adaptive runs. Lines trace a 2-dB range around the predicted amount of binaural unmasking from the Colburn (1977a) model for the 15-cm (dashed black lines) and 1-m (solid gray lines) target. The layout of the spatial configurations of target and masker represented in each panel are shown in the legend. (a) Subject S1. (b) Subject S2. (c) Subject S3.

tive when the target is at (90°, 15 cm) (black circle); in the bottom right panel of Fig. 6(a) [masker at (90°, 15 cm)], the binaural gain is positive when the target is at (90°, 1 m) (gray square).

Overall, target distance has relatively little impact on the binaural component of the spatial release from masking (black and gray symbols are generally comparable within each panel). However, masker distance influences results for all subjects, particularly for the 500-Hz results (Fig. 6) when the masker is located at 90° (right panels). In these configurations, binaural unmasking is smaller when the masker is at 15-cm (lower right panel) than when it is at 1-m (upper right panel).

In general, the binaural contribution to spatial unmasking is larger for the 500-Hz target (Fig. 6) than the 1000-Hz target (Fig. 7). For both target frequencies, the amount of binaural unmasking tends to be largest when the masker is at 0° (left panels in each subplot) and decrease as the masker is displaced laterally (center and right panels in each subplot). Similarly, the change in binaural unmasking with target

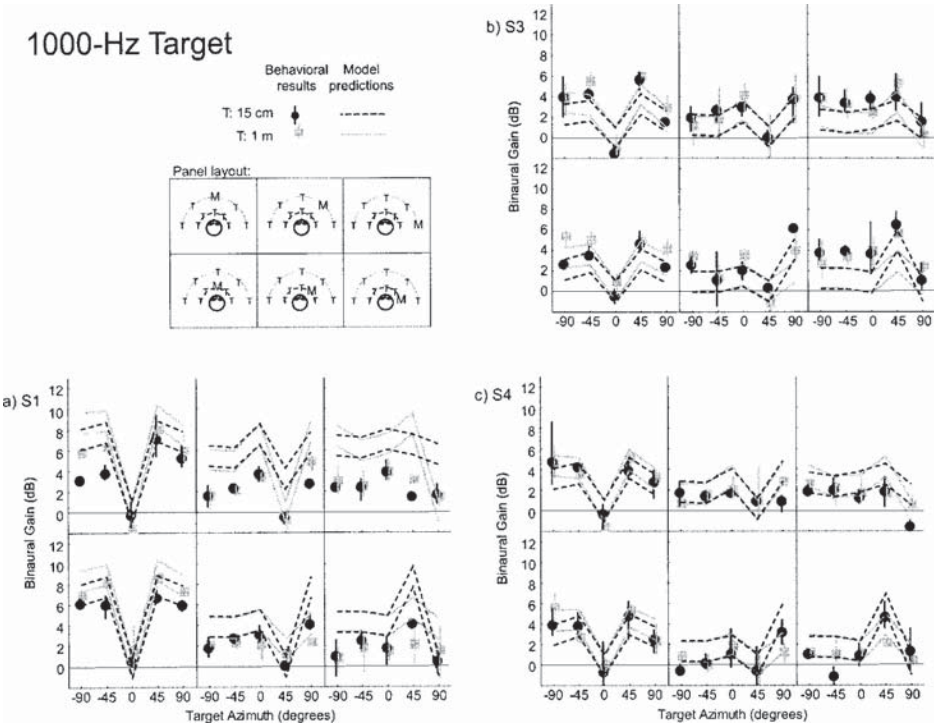

FIG. 7. Estimated binaural contribution to spatial unmasking for the 500-Hz target. See caption for Fig. 6. (a) Subject S1. (b) Subject S3. (c) Subject S4.

angle (i.e., the modulation of binaural gain with target azimuth) is smaller when the masker is laterally displaced (right panels) than when the masker is at 0° (left panels), particularly for the 1000-Hz target (Fig. 7). For instance, looking at the bottom left panel of Fig. 7(a), when the masker is at (0°, 15 cm) the binaural contributions to spatial unmasking for the 1000-Hz target for subject S1 range from 0 to 8 dB depending on the target azimuth. However, when the masker is at (90°, 15 cm) [bottom right panel in Fig. 7(a)], binaural unmasking is roughly constant, independent of target angle (roughly 0–2 dB).

The angular separation of target and masker that leads to the greatest amount of binaural unmasking depends on target frequency. For the 500-Hz target (Fig. 6), binaural unmasking tends to be greatest when target and masker angles differ by about 90° (for example, in the right columns of Fig. 6 where the masker is at 90°, the unmasking is generally greatest when the target is at 0°). However, for the 1000-Hz target (Fig. 7), binaural unmasking tends to be greatest when target and masker angles differ by roughly 45° (in the right columns of Fig. 7 where the masker is at 90°, the amount of unmasking tends to be greatest when the target is at 45°).

C. Discussion

Better-ear factors contribute significantly to spatial unmasking for all of the spatial configurations tested. Better-ear effects are larger at 1000 Hz than 500 Hz and are larger when the target is at 15-cm compared to when the target is at 1-m. The better-ear contribution to spatial unmasking does not always increase monotonically with angular separation of target and masker. In particular, when the masker is at 90°, displacing the target toward the median plane can lead to decreases in the TMR at the better ear, especially if the target and masker are at different distances. This result helps explain why angular separation of target and masker does not always improve detection performance.

Subjects show large differences in their ability to use binaural cues in detection tasks. For subject S1, binaural differences can decrease detection thresholds by as much as 12 dB at 500 Hz [see Fig. 6(a)]; for subject S2 binaural differences provide at most 7 dB of unmasking [Fig. 6(b)]. These intersubject differences in the binaural component of spatial unmasking roughly correlate with differences in BMLDs (Table I); however, intersubject differences in binaural sensitivity for one masker location do not predict results in other spatial configurations. For example, in the 500-Hz conditions when the masker is at 0°, subjects S1 and S3 [left columns in Figs. 6(a) and (c)] have larger binaural components of spatial unmasking than subject S2 [left column in Fig. 6(b)]. However, when the masker is at 90° [right columns of Figs. 6(a)–(c)], all three subjects exhibit essentially the same amount of binaural unmasking. This result suggests that intersubject differences in binaural sensitivity cannot be fully captured with a single ''binaural sensitivity'' parameter at each frequency [the degree to which intersubject differences can be predicted by Colburn's (1977b) model is considered further in Sec. V].

The magnitude of interaural level differences in the masker appears to have a large effect on the amount of binaural masking. For both target frequencies (Figs. 6 and 7), binaural unmasking is greatest when the masker is at 0° (and ITDs and ILDs in the masker are near zero; left columns in each subplot); when the masker is at 45° and 90° (center and right columns in each subplot), the amount of binaural unmasking decreases for the same angular separation of target and masker (i.e., even for roughly the same difference in target and masker ITD). When the masker is off to the side (right columns in the subplots of Figs. 6 and 7), the binaural contribution to spatial unmasking is also smaller when the masker is at 15-cm (when ILDs are very large; bottom right panels) compared to 1-m (when ILDs are smaller; top right panels). These effects are consistent with past reports showing that the BMLD decreases with masker ILD (e.g., see Durlach and Colburn, 1978, p. 433).

In general, the maximum difference in interaural phase difference (IPD) cues for target and masker arises when the ITDs for target and masker differ by one-half the period of the target frequency. For a 500-Hz target, the ITDs in target and masker need to differ by roughly 1 ms to maximize binaural unmasking. For a 1000-Hz target, the ITDs in target and masker need to differ by roughly 500 μs. This explains the dependence of maximal binaural unmasking on target and masker separation and frequency: results in Fig. 5 show that an angular separation of about 90° causes target and masker ITDs to differ by roughly 1 ms (maximizing IPD differences in target and masker for a 500-Hz target) whereas an angular separation of about 45° causes target and masker ITDs to differ by roughly 500 μs.

V. BINAURAL MODEL PREDICTIONS

A. Analysis

Subject-specific predictions of binaural unmasking were calculated using a modified version of the Colburn (1977a, 1977b) model (a description of the current implementation of the model is provided in the Appendix). Predictions depend on six parameters, evaluated at the target frequency: the ITDs and ILDs in both target and masker; the binaural sensitivity of the listener; and the spectrum level of the masker at the more intense ear relative to the absolute, monaural detection threshold in quiet.

The ITDs and ILDs used in the predictions were taken from the analysis of the cues present in the HRTFs. The ITD and ILD in masker were calculated from the values averaged over the ERB filter centered on the target frequency (see Fig. 5). The ITD and ILD in the target were taken directly from the HRTF values at the target frequency (not averaged over the ERB). Binaural sensitivity at each frequency was set to the measured BMLD for each subject and target frequency (Table I). For both the 500- and 1000-Hz targets, the monaural detection threshold (parameter K in the model) was set to 44 dB/Hz.

B. Results

Model predictions are plotted alongside behavioral estimates of the binaural contribution to spatial unmasking in Figs. 6 and 7 (for the 500- and 1000-Hz targets, respectively). In order to be somewhat conservative in identifying

conditions where the model fails to account for behavioral data, parallel lines plot a range of ±1 dB around the actual model predictions. Predictions for the nearby target are shown as dashed black lines; predictions for the far target are shown as solid gray lines.

Model predictions of binaural unmasking are nonnegative for all spatial configurations. Predictions are exactly zero whenever the target and masker are at the same spatial location and positive whenever the target and masker have differences in either their IPDs or ILDs at the target frequency. Thus, in theory, predictions of binaural unmasking are positive whenever the target and masker are at different distances but in the same direction off the median plane because of differences in ILDs in target and masker. However, in practice, predictions are near zero for all configurations when the target and masker are in the same direction for subjects S2, S3, and S4 [Figs. 6(b), 6(c), 7(b), and 7(c)]. Predictions for subject S1 [who has the largest ILDs for 15-cm sources and the largest BMLDs at both frequencies; Figs. 6(a) and 7(a)] are greater than zero for both target frequencies when the target and masker are at different distances but the same (off-median-plane) direction. For instance, in the top center and top right panels of Figs. 6(a) and 7(a) [masker at (45°, 1 m) and (90°, 1 m)], the black dotted lines (predictions for the target at 15 cm) are above zero for all target azimuths, including the target at 90°; in the bottom center and right panels of Figs. 6(a) and 7(a) [masker at (45°, 15 cm) and (90°, 15 cm)], the gray solid lines (predictions for the target at 1 m) are positive for all azimuths.

Binaural unmasking predictions are generally larger at 500 Hz (Fig. 6) than 1000 Hz (Fig. 7). At both frequencies, binaural unmasking varies with angular separation of target and masker; however, the angular separation that maximizes the predicted spatial unmasking depends on frequency. As in the behavioral results, predicted binaural unmasking is greatest when the target and masker are separated in azimuth by 90° for the 500-Hz target (Fig. 6) and 45° for the 1000-Hz target (Fig. 7), corresponding to separations that maximize the differences in target and masker IPD at the target frequency (e.g., in the left column of Fig. 6, when the 500-Hz masker is at 0°, the maximum predicted unmasking, shown by the lines, occurs for targets at +90° and −90°; however, in the left column of Fig. 7, when the 1000-Hz masker is at 0°, the maximum predicted unmasking generally occurs for targets at +45° and −45°).

Also consistent with behavioral results, the maximum predicted amount of binaural unmasking decreases with masker ILD. As a result, the predicted amount of binaural unmasking varies with masker location, systematically decreasing with increasing masker angle and decreasing when the masker is at 15-cm compared to 1-m. For instance, predicted levels of unmasking are generally largest when the masker is at 0° (left columns of Figs. 6 and 7) and decrease as the masker is laterally displaced (center and right columns). Similarly, the amount of unmasking tends to be larger for the top rows of data in Figs. 6 and 7, when the masker is at 1-m, than in the bottom rows of data, when the masker is at 15-cm.

Model predictions capture much of the variation in binaural unmasking; however, there are systematic prediction errors that are large compared to the intrasubject variability. (Note that the standard error in the mean behavioral results is less than or equal to 1 dB as a direct result of the experimental procedure. The error bars in the figure are even more conservative, showing the *range* of thresholds obtained over multiple runs.)

Predictions are first compared to behavioral results for the 500-Hz target (Fig. 6). Predictions for subject S1 agree well with behavioral results when the masker is at (0°, 15 cm) [bottom left panel of Fig. 6(a)] and reasonably well for three other masker locations [(45°, 15 cm), (90°, 15 cm), and (90°, 1 m); bottom center, bottom right, and top right panels of Fig. 6(a), respectively]. However, S1 predictions tend to overestimate binaural unmasking for two masker locations [(0°, 1 m) and (45°, 1 m); top left and top center panels of Fig. 6(a)]. For subject S2, predictions match behavioral results reasonably well when the masker is at 0° [see the top left and bottom left panels of Fig. 6(b)], independent of masker distance (although there are isolated data points for which the model overestimates binaural unmasking), but systematically underestimate binaural unmasking when the masker is at 45° and 90° for both masker distances [see center and right panels of Fig. 6(b), where symbols fall above lines]. Results for subject S3 are similar to those of subject S2: predictions are in good agreement with measurements when the masker is in the median plane [left panels of Fig. 6(c)] but underestimate binaural unmasking when the masker is laterally displaced [center and right panels of Fig. 6(c)].

Focusing on the 1000-Hz results (Fig. 7), subject S1 predictions generally overestimate binaural unmasking (in all panels in Fig. 7(a), symbols fall below lines). For subject S3, predictions generally underestimate binaural unmasking, except when the masker is at (45°, 1 m), where predictions and measurements are reasonably close [agreement between the measured data points and the prediction lines is good only for the top center panel of Fig. 7(b); for all other panels, symbols fall above lines]. Finally, predictions for subject S4 either fit reasonably well or underestimate binaural unmasking when the masker is at 0° [left panels of Fig. 7(c)] but overestimate binaural unmasking when the masker is at 45° or 90°, independent of masker distance [see center and right panels of Fig. 7(c), where symbols fall below lines].

Overall, predictions and behavioral results are in better agreement when the masker is in the median plane than when the masker is at 45° or 90° and for the 500-Hz data compared to the 1000-Hz data.

C. Discussion

The Colburn model assumes that a single value representing binaural sensitivity at a particular frequency can account for intersubject differences in binaural unmasking. This binaural sensitivity parameter was set from BMLD measures taken with a diotic masker and target that was either diotic (NoSo) or inverted at one ear to produce an interaural phase difference of π (NoSπ). These conditions are most analogous to the spatial configurations in which the masker is directly in front of the listener (and the masker is

essentially diotic). For most of the configurations with the masker at 0°, model predictions agree well with observed results. In contrast, larger discrepancies between the modeled and measured results arise when the masker is at 45° and 90° (conditions in which there are significant ILDs in the masker).

While there are some conditions in which the model predictions consistently over- or underestimate binaural unmasking [e.g., results for subject S1 at 1000 Hz in Fig. 7(a) or for subject S3 at 1000 Hz in Fig. 7(b)], there are other conditions for which changing the single subject-specific ''binaural sensitivity'' of the model cannot account for discrepancies between the model predictions and the measurements [e.g., results for subject S2 at 500 Hz in Fig. 6(b) or for subject S4 at 1000 Hz in Fig. 7(c)].

The current results suggest that subjects differ not only in their overall sensitivity to binaural differences, but also in the dependence of binaural sensitivity on the interaural parameters in masker and/or target. In particular, binaural sensitivity appears to depend on the interaural level difference in the masker differently for different subjects. As a result, individualized model prediction errors are generally larger when there are large ILDs in the masker than when the masker has near-zero ILD. While the Colburn model has been tested (and shown to predict results relatively well) in many studies in which target and masker vary in their interaural phase parameters, there are few studies that manipulate the target and masker ILD. These results suggest the need for additional behavioral and theoretical studies of the effects of ILD in binaural detection tasks.

Even though there are specific conditions for which predictions fail to account for the results for a particular subject, the model captures many of the general patterns in results, including the tendency for binaural unmasking to decrease as the ILD in the masker increases and how the amount of binaural unmasking depends on the angular separation of target and masker and the frequency of the target.

VI. GENERAL DISCUSSION

The current study is unique in measuring how tone detection thresholds are affected by target and masker location when sources are very close to the listener. Results show that for sources very close to the listener, small changes in source location can lead to large changes in detection threshold. These large changes arise from changes in both the TMR (affecting the better-ear contribution to spatial unmasking) and ILDs (affecting the binaural contribution to spatial unmasking).

The current results demonstrate how the relative importance of better-ear and binaural contributions to spatial unmasking change with target and masker location, including source distance (in contrast to previous studies that considered only angular separation of relatively distant sources). The relative importance of better-ear contributions to spatial unmasking increases as masker distance decreases, probably because of increases in the ILD in the masker, which reduce the amount of binaural unmasking. The better-ear contribution also increases as target distance decreases, primarily because the TMR changes more rapidly with target angle when the target is near the listener. The relative importance of the better-ear contribution to spatial unmasking increases with target frequency, both because the absolute magnitude of better-ear factors increases and because the binaural contribution to unmasking decreases. For a 500-Hz target, binaural and better-ear factors are roughly equally important when the masker is in the median plane. However, better-ear factors become relatively more important as the masker is displaced laterally, in part because the amount of binaural spatial unmasking decreases with masker ILD. This trend, which is predicted by the Colburn model, helps to explain large differences in the amount of spatial unmasking observed in previous studies (e.g., Ebata *et al.*, 1968; Gatehouse, 1987; Santon, 1987). Specifically, more spatial unmasking arises when the masker is positioned in front of the listener and the target location is varied (leading to near-zero ILDs in the masker) than when the target is fixed in location and the angle of masker is varied (leading to progressively larger ILDs in the masker with spatial separation of target and masker).

Binaural processing contributes up to 10 dB to spatial unmasking for the spatial configurations tested. In theory, differences in target and masker distance cause differences in target and masker ILD when the sources are off the median plane, leading to binaural unmasking. However, in the current study evidence of binaural unmasking resulting from differences in target and masker distance was observed only for Subject S1, who had both the largest BMLDs and the largest ILDs of the four subjects in the study.

Although monaural detection thresholds were not directly measured in the current study, binaural performance is always better than or equal to the performance predicted by analysis of the TMR at the better ear. Thus, the current study does not help to explain results suggesting that binaural performance sometimes falls below monaural performance using the better ear alone, particularly for configurations with large ILDs (Bronkhorst and Plomp, 1988; Shinn-Cunningham *et al.*, 2001). One important distinction between the current study and these previous reports is that the current study measured tone detection for relatively low-frequency tones, whereas both of the previously cited studies measured speech intelligibility, a suprathreshold task that emphasizes information at higher frequencies. Further studies are necessary to help determine when binaural stimulation may actually degrade performance compared to monaural, better-ear performance.

Intersubject differences in the amount of spatial unmasking are large and arise from individual differences in (1) HRTFs, (2) overall binaural sensitivity, and (3) the way in which binaural sensitivity varies with spatial configuration of target and masker. The Colburn (1977b) model of binaural processing predicts overall trends in behavioral measures of binaural unmasking, but fails to capture subject-specific variations in performance. The spatial configurations for which model predictions are least accurate are the positions for which large ILDs arise in masker and/or target, conditions that have not been extensively tested in previous studies. The current results suggest that the Colburn model must be modified so that subject differences in binaural sensitivity

vary not only in overall magnitude but as a function of the interaural differences in the masker.

While predictions from the Colburn model (taking into account differences in the stimuli presented to the individual subjects as well as individual differences in binaural sensitivity) cannot account for some small but significant intersubject differences in spatial unmasking, rough predictions of the amount of spatial unmasking capture most of the observed changes in detection threshold with spatial configuration. For instance, generic acoustic models of HRTFs (e.g., KEMAR measurements or spherical-head model predictions) combined with predictions of binaural unmasking using "average" model parameters should produce predictions that fall within the range of behavior observed across a population of subjects.

VII. CONCLUSIONS

(1) Acoustic cues (particularly TMR and ILD) vary dramatically with source distance and direction for nearby sources. Therefore, when source distance varies, the effect of source location on both the better-ear and binaural contributions to spatial unmasking is complex.
(2) For nearby sources, the better-ear contribution to pure-tone spatial unmasking can be very large (as much as 25 dB) compared to conditions where sources are relatively far from the listener.
(3) The binaural contribution to spatial unmasking decreases with increasing masker ILD. As a result, the binaural contribution to spatial unmasking is smaller for lateral sources very near the head than for more distant sources at the same lateral angle relative to the listener.
(4) Intersubject differences in spatial unmasking are larger for nearby sources than for far sources, in part because there are larger acoustic differences in HRTFs for nearby sources compared to more distant sources. However, there also are subject-specific differences both in binaural sensitivity and on how ILDs influence binaural sensitivity.
(5) Predictions based on Colburn's analysis (1977b) show the correct general trends in binaural detection for both near and far sources, but cannot account for small, but consistent, subject-specific differences in performance, particularly when large ILDs are present in the masker.

ACKNOWLEDGMENTS

This work was supported in part by AFOSR Grant No. F49620-98-1-0108 and the Alfred P. Sloan Foundation. Portions of this work were presented at the Spring 2000 meeting of the Acoustical Society of America. H. Steven Colburn provided valuable input throughout this work, including help in putting the results in appropriate context. Les Bernstein, Adelbert Bronkhorst, and an anonymous reviewer provided valuable criticism and comments on a previous draft of this paper.

APPENDIX: BINAURAL MODELING

A modified version of the model presented in Colburn (1977b) was used to predict the amount of binaural unmasking, defined as the difference in detection thresholds when target and masker are at the same spatial location and when they are in different locations. The predicted amount of binaural unmasking for a target at frequency f_0 is computed as

$$s(f_0,\alpha_T,\phi_T,\alpha_M,\phi_M,\mathrm{BMLD},K) = \sqrt{\max\left(1,\frac{\alpha_T^4}{\alpha_M^4}\right)+(2\cdot 10^{\mathrm{BMLD}/10}-1)R(\alpha_M,10^{K/10})\frac{F^2(\phi_M,f_0)}{16}\left(1+\frac{\alpha_T^2}{\alpha_M^2}-2\frac{\alpha_T}{\alpha_M}\cos(\phi_M-\phi_T)\right)^2}, \quad \text{(A1)}$$

where $\alpha_T=10^{\mathrm{ILD}-T/20}$; $\alpha_M=10^{\mathrm{ILD}-M/20}$; ILD$-T$ and ILD$-M$ are the interaural level differences in target and masker (respectively) in dB; ϕ_T and ϕ_M are the IPDs of target and masker (respectively) in radians; BMLD is the (subject-specific) binaural masking level difference in dB; K is the level of masker relative to absolute detection threshold in quiet, in dB; and the functions F^2 and R are defined below (all evaluated at the target frequency).

Function F^2 represents the extent to which phase shifts in masker cannot be compensated by internal time delays. This function is given by

$$F^2(\phi_M,f_0)=\frac{\Sigma_{k=-1000}^{1000}p(\phi_M/2\pi f_0+k/f_0,f_0)\exp\{-G^2(f_0)[1-\gamma(\phi_M/2\pi f_0+k/f_0)]\}}{\Sigma_{k=-1000}^{1000}p(k/f_0,f_0)\exp\{-G^2(f_0)[1-\gamma(k/f_0)]\}}, \quad \text{(A2)}$$

where $p(\tau,f)$ represents the relative number of interaural coincidence detectors (i.e., neurons in the medial superior olive) tuned to ITD τ and frequency f; $G(f)$ represents the synchrony of firings of the auditory nerve at frequency f (squared to account for the sharpening of synchrony in the cochlear nucleus); and $\gamma(\tau)$ is the envelope of the autocorrelation function of the auditory nerve fiber impulse response at autocorrelation delay τ. In the current realization of the model, function $p(\tau,f)$ was modified to allow for a frequency dependence in the distribution of interaural coincidence detectors (as suggested by Stern and Shear, 1996), using

$$p(\tau,f_0)=\begin{cases} C(e^{-2\pi k_l|0.2|}-e^{-2\pi k_h|0.2|})/0.2, & |\tau|<0.2 \text{ ms},\\ C(e^{-2\pi k_l|0.2|}-e^{-2\pi k_h|0.2|})/|\tau|, & |\tau|\geqslant 0.2 \text{ ms},\end{cases}$$

$$k_h=3\times 10^6,$$

$$k_l\begin{cases} 0.1(f_0 10^{-3})^{1.1}, & f_0\leqslant 1200 \text{ Hz},\\ 0.1(1200\times 10^{-3})^{1.1}, & f_0>1200 \text{ Hz},\end{cases} \tag{A3}$$

$$C=\begin{cases} 0.1534, & f_0=500 \text{ Hz},\\ 0.2000, & f_0=1000 \text{ Hz}.\end{cases}$$

$G(f)$ is given by

$$G(f_0)=\begin{cases} \sqrt{10}, & f_0\leqslant 800 \text{ Hz},\\ \sqrt{10}\dfrac{800}{f_0}, & f_0>800 \text{ Hz}.\end{cases} \tag{A4}$$

$\gamma(\tau)$ is given by

$$\gamma(\tau)=\begin{cases} 2.359\times 10^{-4}+1.5207\times 10^8\tau^4-1.764\times 10^4\tau^2 & \\ \qquad +0.993, & |\tau|\leqslant 0.006,\\ -97.3236|\tau|+1.139, & |\tau|>0.006,\end{cases} \tag{A5}$$

where τ is in milliseconds.

Finally, function $R(\alpha,K)$ characterizes the decrease in the number of activated auditory nerve fibers in the ear receiving the less intense signal as a function of masker ILD. The current implementation uses a modified version of Eq. (35) from Colburn (1977b):

$$R(\alpha_n)=\begin{cases} \left(\dfrac{10\log_{10}\alpha_n^{-2}K}{40}\right)^2, & \alpha_n^{-2}K\leqslant 10^4,\\ 1, & \alpha_n^{-2}K>10^4,\end{cases} \tag{A6}$$

where K is the ratio of the spectrum level at the more intense ear to the detection threshold in quiet. This implementation of the model assumes that the auditory nerve fibers at each target frequency have thresholds uniformly distributed (on a dB scale) over a 40-dB range above the absolute detection threshold for that frequency.

[1]System identification using a MLS depends on circular convolution techniques. Theoretically, the approach requires the MLS to be concatenated with itself and presented an infinite number of times to ensure that the system is in its steady-state response prior to measuring the response (see Vanderkooy, 1994). The resulting estimated system response is a time-aliased version of the true system response. In the current measures, the MLS was presented twice and the response to the second repetition was recorded. Given the length of the MLS used, the room characteristics of and ambient noise in the environment in which we were measuring, and the noise in our measurement system, the steady-state response can be approximated with only two repetitions of the MLS and no significant time aliasing is present in our measurements.

[2]Note that this analysis assumes that detection performance depends only on the target-to-masker ratio or TMR and is independent of the overall masker level, an assumption that is not valid if the masker is near absolute threshold or at very high presentation levels. For instance, imagine two masker locations so distant from the listener that the masker is inaudible. These masker locations would produce identical signal detection thresholds if the experiment were performed with the distal stimulus intensity fixed; however, our technique might adjust the masker by different amounts for these two masker locations in order to achieve a fixed proximal stimulus level at the ear of the listener, producing two different estimates of spatial unmasking. While holding the distal masker intensity fixed may seem more natural and intuitive than holding the proximal stimulus level constant, the overall presentation level of the masker would span an extraordinarily large range in the current experiments because the masker distance varied between 15 cm and 1 m in addition to varying in direction. Therefore, we elected to fix the proximal masker intensity.

Bronkhorst, A. W., and Plomp, R. (**1988**). "The effect of head-induced interaural time and level differences on speech intelligibility in noise," J. Acoust. Soc. Am. **83**, 1508–1516.

Brungart, D. S., and Rabinowitz, W. M. (**1999**). "Auditory localization of nearby sources. I. Head-related transfer functions," J. Acoust. Soc. Am. **106**, 1465–1479.

Colburn, H. S. (**1977a**). "Theory of binaural interaction based on auditory-nerve data. II: Detection of tones in noise," J. Acoust. Soc. Am. **61**, 525–533.

Colburn, H. S. (**1977b**). "Theory of binaural interaction based on auditory-nerve data. II: Detection of tones in noise. Supplementary material," J. Acoust. Soc. Am. AIP document no. PAPS JASMA-61-525-98.

Doll, T. J., and Hanna, T. E. (**1995**). "Spatial and spectral release from masking in three-dimensional auditory displays," Hum. Factors **37**, 341–355.

Duda, R. O., and Martens, W. L. (**1998**). "Range dependence of the response of a spherical head model," J. Acoust. Soc. Am. **104**, 3048–3058.

Durlach, N. I., and Colburn, H. S. (**1978**). "Binaural phenomena," in *Handbook of Perception*, edited by E. C. Carterette and M. P. Friedman (Academic, New York), pp. 365–466.

Ebata, M., Sone, T., and Nimura, T. (**1968**). "Improvement of hearing ability by directional information," J. Acoust. Soc. Am. **43**, 289–297.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.

Gatehouse, R. W. (**1987**). "Further research on free-field masking," J. Acoust. Soc. Am. Suppl. 1 **82**, S108.

Good, M. D., Gilkey, R. H., and Ball, J. M. (**1997**). "The relation between detection in noise and localization in noise in the free field," in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. Gilkey and T. Anderson (Erlbaum, New York), pp. 349–376.

Kidd, Jr., G., Mason, C. R., Rohtla, T. L., and Deliwala, P. S. (**1998**). "Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns," J. Acoust. Soc. Am. **104**, 422–431.

Levitt, H. (**1971**). "Transformed up-down methods in psychophysics," J. Acoust. Soc. Am. **49**, 467–477.

Moore, B. C. J. (**1997**). *An Introduction to the Psychology of Hearing*, 4th ed. (Academic, San Diego).

Saberi, K., Dostal, L., Sadralodabai, T., Bull, V., and Perrott, D. R. (**1991**). "Free-field release from masking," J. Acoust. Soc. Am. **90**, 1355–1370.

Santon, F. (**1987**). "Detection d'un son pur dans un bruit masquant suivant l'angle d'incidence du bruit. Relation avec le seuil de reception de la parole," Acustica **63**, 222–230.

Shinn-Cunningham, B. G., Santarelli, S., and Kopco, N. (**2000**). "Tori of confusion: Binaural localization cues for sources within reach of a listener," J. Acoust. Soc. Am. **107**, 1627–1636.

Shinn-Cunningham, B. G., Schickler, J., Kopco, N., and Litovsky, R. Y. (**2001**). "Spatial unmasking of nearby speech sources in a simulated anechoic environment," J. Acoust. Soc. Am. **110**, 1118–1129.

Stern, R. M., and Shear, G. D. (**1996**). "Lateralization and detection of low-frequency binaural stimuli: Effects of distribution of internal delay," J. Acoust. Soc. Am. **100**, 2278–2288.

Vanderkooy, J. (**1994**). "Aspects of MLS measuring systems," J. Audio Eng. Soc. **42**, 219–231.

In: Auditory signal processing: Physiology, psychoacoustics, and models. (Pressnitzer, D., de Cheveigné, A, McAdams, S., and Collet, L., eds), pp 327-333, Springer, New York. (Proc. International Symposium on Hearing, Dourdan, France, Aug. 24-29, 2003)

A cat's cocktail party: Psychophysical, neurophysiological, and computational studies of spatial release from masking

Courtney C. Lane[1], Norbert Kopco[2], Bertrand Delgutte[1], Barbara G. Shinn-Cunningham[2], and H. Steven Colburn[2]

1 Eaton-Peabody Laboratory, Massachusetts Eye and Ear Infirmary, Boston, MA, USA {court, bard}@epl.meei.harvard.edu

2 Hearing Research Center, Boston University, Boston, MA, USA {kopco, shinn, colburn}@bu.edu

1 Introduction

Masked thresholds can improve substantially when a signal is spatially separated from a noise masker (Saberi et al. 1991). This phenomenon, termed "spatial release from masking" (SRM), may contribute to the cocktail party effect, in which a listener can hear a talker in a noisy environment. The purpose of this study is to explore the underlying neural mechanisms of SRM.

Previous psychophysical studies (Good, Gilkey, and Ball 1997) have shown that for high-frequency stimuli, SRM was due primarily to energetic effects related to the head shadow, but for low-frequency stimuli, both binaural processing (presumably ITD processing) and energetic effects contributed to SRM. The relative contributions of these two factors were not studied for broadband stimuli.

Previous physiology studies have identified possible neural substrates for both the energetic and ITD-processing components of SRM. For the energetic component, our group has shown that some inferior colliculus units, "SNR units," have masked thresholds that are predicted by the signal-to-noise ratio (SNR) in a narrowband filter centered at the unit's CF (Litovsky et al. 2001). For the ITD component, a series of studies (e.g. Jiang, McAlpine, and Palmer 1997) shows that ITD-sensitive units can exploit the differences between the interaural phase difference (IPD) of a tone and masker to improve the neural population masked thresholds. These studies did not describe how the units' masked thresholds change when a broadband signal and masker are placed at different azimuths.

Here, we examine the contributions of energetic effects and binaural processing for broadband and low-frequency SRM using psychophysical experiments and an idealized population of SNR units. We also show that a population of ITD-sensitive units in the auditory midbrain exhibits a correlate of SRM. Finally, a model of ITD-sensitive units reveals that the signal's temporal envelope influences the single-unit masked thresholds.

2 Psychophysics and modeling of SRM in humans

2.1 Methods

SRM was measured for three female and two male normal-hearing human subjects using lowpass and broadband stimuli. Azimuth was simulated using non-individualized head-related transfer functions (Brown 2000). Stimuli consisted of a 200-ms 40-Hz chirp train (broadband: 300-12,000 Hz; lowpass: 200-1500 Hz) masked by noise (broadband: 200-14,000 Hz, lowpass: 200-2000 Hz). The spectrum-level for the signal was fixed at 14 dB re 20 μPa/√Hz (56 dB SPL for the broadband signal). The masker level was adaptively varied using a 3-down, 1-up procedure to estimate the signal-to-noise ratio (SNR) yielding 79.4% correct detection performance. Stimuli were delivered via insert earphones to subjects in a sound-treated booth.

Inspired by the SNR units described above, predictions from a simple, "single-best-filter" model were used to evaluate if the SNR in the best narrow-frequency band can explain how masked threshold varies with signal and noise locations. The model analyzes SNR as a function of frequency, but does not allow for any across-frequency integration of information or any binaural processing. The model consists of a bank of 60 log-spaced gammatone filters (Johannesma 1972) for each ear. For each spatial configuration, the root-mean-squared energy at the output of every filter is separately computed for the signal and noise. The model assumes that the filter with the largest SNR (over the set of 120) determines threshold. The only free parameter in the model, the SNR yielding 79.4% correct performance, was fit to match the measured threshold when signal and noise were at the same location.

2.2 Results

Figure 1 shows measured (solid lines) and predicted (broken lines) thresholds as a function of noise azimuth for three signal azimuths (arrows). Two sets of model predictions are shown. Dash-dot lines show both lowpass and broadband predictions generated jointly for the model parameter fit to the broadband threshold measured with signal and masker co-located. Dotted lines show lowpass predictions generated with the model parameter fit to the measured lowpass threshold separately. Overall, performance is better for broadband (BB) stimuli than for lowpass (LP) stimuli (BB thresholds are always lower than LP). Further, the amount of SRM, the improvement in threshold SNR compared to the thresholds when signal and noise are co-located, is larger for broadband than lowpass stimuli (30 dB and 12 dB, respectively).

When the model parameter is fit separately for broadband and lowpass stimuli, predictions are relatively close to observed thresholds although lowpass predictions consistently underestimate SRM. These results suggest that for the chirp-train signals used, 1) the main factor influencing SRM for both lowpass and broadband stimuli is the change in SNR in narrow frequency bands, and 2) binaural processing increases SRM for lowpass, but not broadband stimuli.

When the same threshold SNR parameter is used to predict broadband and lowpass results (dash-dot lines), predicted thresholds are equal when signal and

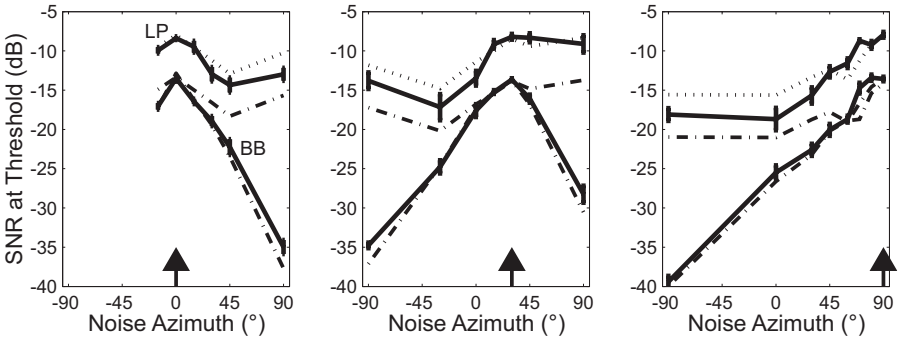

Fig. 1. SRM for human subjects for broadband (BB) and lowpass (LP) stimuli. Measured (subject mean and standard error) and predicted thresholds as a function of noise azimuth for three signal azimuths (arrows). Dash-dot line: lowpass and broadband model fit with same parameter; dotted line: lowpass data fit separately.

noise are co-located, regardless of stimulus bandwidth (because the SNR is constant across frequency when signal and noise are co-located). However, measured performance is always worse for the lowpass stimuli compared to the broadband stimuli. This result suggests that the listener integrates information across frequency, leading to better performance for broadband stimuli.

3 Neural correlates of SRM in the cat auditory midbrain

As shown above, the single-best-filter model underestimates the SRM for low frequencies. Here, thresholds for a population of ITD-sensitive neurons are measured to determine if these units can account for the difference between the single-best-filter model and behavioral thresholds.

3.1 Methods

Responses of single units in the anesthetized cat inferior colliculus were recorded using methods similar to those described in Litovsky and Delgutte (2002). The signal was a 40-Hz, 200-msec chirp train presented in continuous noise; both signal and noise contained energy from 300 Hz to 30 kHz. The chirp train had roughly the same envelope as the one used in the broadband psychophysical experiments. The signal level was fixed near 40 dB SPL, and the noise level was raised to mask the signal response. Results are reported for 22 ITD-sensitive units with characteristic frequencies (CFs) between 200 and 1200 Hz.

3.2 Results

Figure 2A shows the temporal response pattern for a typical ITD-sensitive unit as a function of noise level for the signal in noise (first 200 msec) and the noise alone (second 200 msec). The signal and noise were both placed at +90° (contralateral to the recording site). At low noise levels, the unit produces a synchronized response to the 40-Hz chirp train. As the noise level increases, the response to the signal is

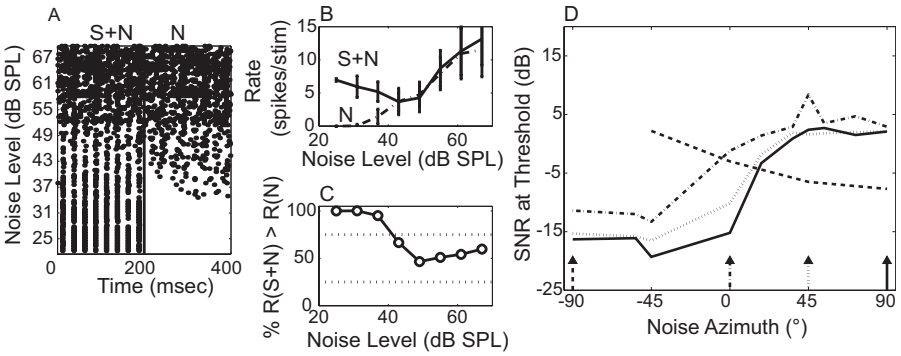

Fig. 2. A: Single-unit response pattern for signal in noise (S+N, 0-200 msec) and noise alone (N, 200-400 msec) for signal and noise at 90°. Signal level is 43 dB SPL. B: Rate-level functions for S+N and N from A. C: Percent of stimulus presentations that have more spikes for S+N compared to N. Threshold is the SNR at 75% or 25% (dotted lines). D: Same unit's masked thresholds as a function of noise azimuth for four signal azimuths (arrows indicate signal azimuth, arrow tail indicates corresponding threshold curve).

overwhelmed by the response to the noise (A, B). For this unit, +90° is a favorable azimuth so both the signal and the noise excite the unit. When placed at an unfavorable azimuth, the signal can suppress the noise response or vice versa.

Threshold is defined for single units as the SNR at which the signal can be detected through a rate increase or decrease for 75% of the stimulus repetitions (75% and 25% lines in Fig. 2C). Thresholds for this unit are shown in D as a function of noise azimuth for four signal azimuths. For three of the signal azimuths (-90°, 45°, and 90°), moving the noise away from the signal can improve thresholds by more than 15 dB. However, when the signal is at 0°, thresholds become slightly worse as the noise moves from the midline to the contralateral (positive azimuth) side. In other words, although some SRM is seen for some signal azimuths, no direct correlate of SRM can be seen in this, or any other, individual unit's responses for all signal and noise configurations.

A simple population threshold is constructed based on the same principle as the single-best-filter model (Section 2). For each signal and noise configuration, the population threshold is the best single-unit threshold in our sample of ITD-sensitive

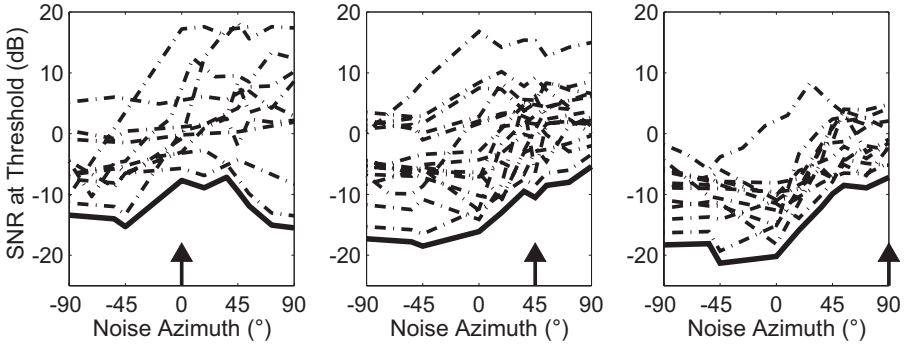

Fig. 3. Neural population thresholds for three signal azimuths (arrow). Dash-dot lines: single unit thresholds; solid lines: population thresholds (offset by 2 dB).

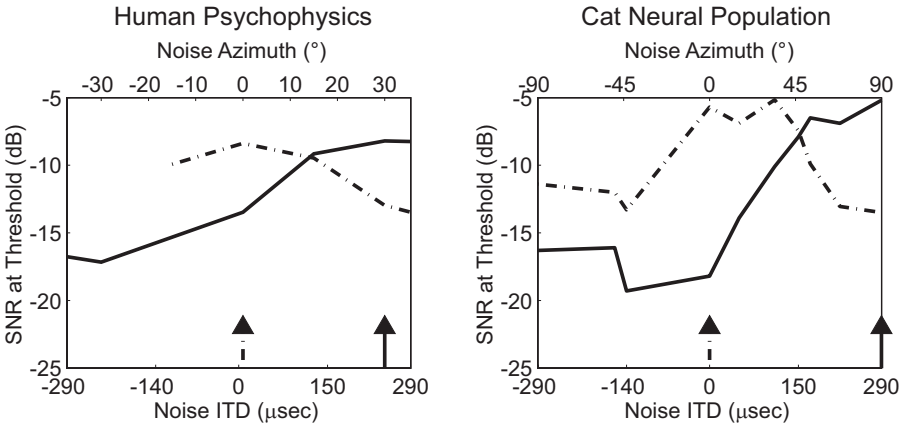

Fig. 4. Human psychophysical thresholds (left) and cat neural population thresholds (right) for two signal azimuths (arrows indicate signal azimuth, arrow tail indicates corresponding threshold curve) as a function of noise ITD (lower axis) and azimuth (upper axis).

units. Figure 3 shows the population thresholds (solid lines) as a function of noise azimuth for three signal azimuths (arrows). Unlike single unit thresholds (dot-dash), the population thresholds show SRM in that thresholds improve when the signal and noise are separated.

Figure 4 compares the low-pass human psychophysical thresholds (left) to the cat neural population thresholds (right). In order to compare the two thresholds despite the difference in species headsize, the axes are matched for noise ITD (lower axis) rather than noise azimuth (upper axis). The neural population thresholds are similar to the human behavioral thresholds, indicating that these ITD-sensitive units could provide a neural substrate for the binaural component of SRM.

3.3 Neural modeling of single-unit thresholds

Because our population consists of ITD-sensitive units, we attempted to model the unit responses using an interaural cross-correlator model similar to Colburn (1977). Figure 5A shows the thresholds for five units for which we measured thresholds for the signal at their best azimuths (+90°, squares) and their worst azimuths (-90°, circles). The noise was placed at the ear opposite the signal. For the data, the best-azimuth thresholds are better or equal to the worst-azimuth thresholds. In contrast, the cross-correlator model predicts that the worst-azimuth thresholds are better (Fig. 5B) because the largest change in interaural correlation occurs when the signal decreases the overall correlation. The cross-correlator, although able to predict the noise-alone response, failed to predict the response to the signal (not shown). The primary difference between the chirp-train signal and the noise is that the signal has a strong 40-Hz amplitude modulation while the noise envelope is relatively flat. Because many units in the IC have enhanced responses to modulated stimuli (Krishna and Semple 2000), we added an envelope processor that changes the rate response in proportion to the energy in the 40-Hz Fourier component of the cross-correlator's output. With envelope processing (Fig. 5C), best-azimuth thresholds are

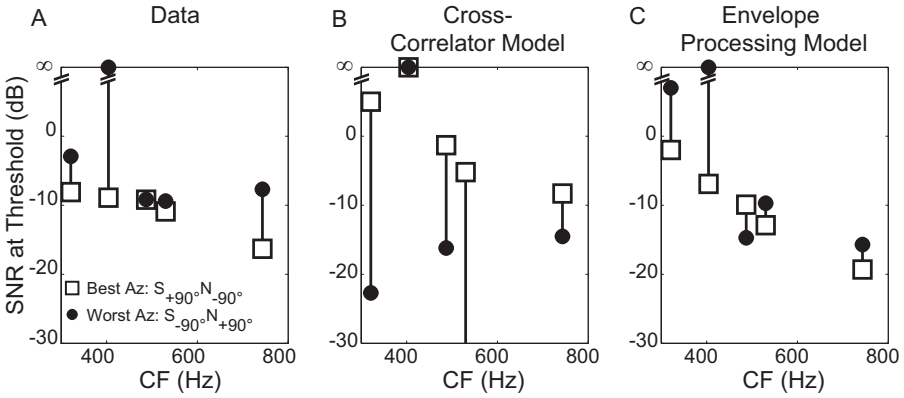

Fig. 5. A: Masked thresholds for 5 units. Best-azimuth thresholds (squares): signal at +90°, noise at -90°; worst-azimuth thresholds (circles): signal at -90°, noise at +90°. B,C: As in A for cross-correlator model (B) and cross-correlator model with envelope processor (C).

about the same or better than worst-azimuth thresholds, consistent with the data, because the envelope processor only changes the responses for favorable azimuths. These results suggest that 1) a traditional cross-correlator model cannot account for neural responses in the IC, 2) the temporal envelope can affect the detectability of signals in inferior colliculus neural responses, and 3) envelope processing is necessary to predict which units are best for signal detection (discussed below).

4 Discussion

Human listeners exhibit a large amount of SRM for both broadband and lowpass 40-Hz chirp-train signals. For broadband stimuli, the SNR in a single high-frequency filter predicts the amount of SRM, indicating high-frequency narrow-band energetic changes determine the SRM. SNR units, which have thresholds that are predicted by the SNR in a narrowband filter, could detect these changes.

For the lowpass condition, the single-best-filter model predicts some SRM, but underestimates the total amount by several dB. A correlate of the lowpass SRM is evident in the population response of ITD-sensitive units in the IC. It is possible, then, that there are two populations of neurons that can give SRM at low frequencies: an ITD-sensitive population and an SNR-unit population. When a listener is able to use the ITD-sensitive population, thresholds should improve by a few dB. When this population cannot be used (such as when the signal and masker are co-located or when listening monaurally), the SNR-unit population would determine performance, resulting in worse masked thresholds for some spatial configurations. These two hypothesized neural populations may respond differently to different stresses. For example, because the SNR population response depends on a neural population with narrow tuning and a wide range of CFs, relying on this population might be especially difficult for listeners with hearing impairment.

The envelope-processing model predicts that different ITD-sensitive populations, in either the left IC or the right IC, will dominate signal detection performance for different stimuli. The best single-unit thresholds for both the data

and the envelope-processing model occur when the chirp-train signal is positioned at a unit's best azimuth. Thus, for modulated signals, the IC contralateral to the signal yields better thresholds than the ipsilateral IC. However, for unmodulated signals, the model predicts that the best thresholds occur for the signal placed at the unit's worst azimuth. This prediction is consistent with previous studies (e.g. Jiang, McAlpine, and Palmer 1997) showing that the best single-unit thresholds for tones in noise occurred when the tone had an unfavorable IPD. Therefore, different ICs seem to be used for signal detection depending on the signal envelope.

Finally, human broadband thresholds are better than lowpass thresholds for all spatial configurations. Because this improvement is evident for co-located signals and maskers, the auditory system seems to integrate information across frequency. Because units in the IC are relatively narrowly tuned, auditory centers above the IC are also likely to be involved in the detection of broadband signals.

In summary, SRM seems to depend on binaural and energetic cues, which may be processed by separate neural populations. Neural processing related to SRM can be observed in the auditory midbrain, but centers higher than the midbrain also seem necessary for the integration of information across frequency.

References

Brown, T. J. (2000). "Characterization of acoustic head-related transfer functions for nearby sources," unpublished M.Eng. thesis. Electrical Engineering and Computer Science, MIT, Cambridge, MA.

Colburn, H. S. (1977) Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise. J. Acoust. Soc. Am. 61, 525-533.

Good, M.D., Gilkey, R.H., and Ball, J.M. (1997) The relation between detection in noise and localization in noise in the free field. In R.H. Gilkey and T.R. Anderson (Eds), Binaural and Spatial Hearing in Real and Virtual Environments. Lawrence Erlbaum Associates, Mahwah, N.J, pp 349–376.

Jiang, D., McAlpine, D., and Palmer, A.R. (1997) Detectability index measures of binaural masking level difference across populations of inferior colliculus neurons. J. Neurosci. 17, 9331-9339.

Johannesma, P.I.M. (1972) The pre-response stimulus ensemble of neurons in the cochlear nucleus. In: B.L. Cardozo, E. de Boer, and R. Plomp (Eds.), IPO Symposium on Hearing Theory. IPO, Eindhoven, The Netherlands, pp. 58-69.

Krishna, B.S. and Semple, M.N. (2000) Auditory temporal processing: responses to sinusoidally amplitude-modulated tones in the inferior colliculus. J. Neurophysiol. 84, 255-73.

Litovsky, R.Y. and Delgutte, B. (2002) Neural correlates of the precedence effect in the inferior colliculus: Effect of localization cues. J. Neurophysiol. 87, 976-994.

Litovsky, R.Y., Lane, C.C., Atencio, C., and Delgutte, B. (2001) Physiological measures of the precedence effect and spatial release from masking in the cat inferior colliculus. In: D.J. Breebaart, A.J.M. Houtsma, A. Kohlrausch, V.F. Prijs, and R. Schoonhoven (Eds). Physiological and Psychophysical Bases of Auditory Function. Shaker, Maastricht, pp. 221-228.

Saberi, K., Dostal, L., Sadralodabai,T., Bull, V., and Perrott, D.R. (1991) Free-field release from masking. J. Acoust. Soc. Am. 90, 1355-1370.

ForumAcusticum 2005

In: Forum Acusticum 2005 : Proceedings : 29 Aug - 2 Sep 2005, Budapest, Hungary. Budapest : OPAKFI, 2005. p. 1607-1612.

Across-frequency integration in spatial release from masking

Norbert Kopčo
Department of Cybernetics and AI, Technical University, Košice, Slovakia and
Hearing Research Center, Boston University, **kopco@bu.edu**

Spatial separation of a target (T) stimulus from a masker (M) often improves detectability of the target, a phenomenon known as the spatial release from masking (SRM). When the masker is a noise, two main factors contribute to SRM: changes in the target-to-masker ratio dominate the performance at high frequencies, while binaural processing dominates at low frequencies. Previous neurophysiological studies (e.g., Lane et al., ISH 2003) suggested that, at the level of inferior colliculus, the SRM of broadband stimuli is determined by a single unit – the one that is most sensitive in the given T/M spatial configuration. Based on this observation, Lane et al. proposed a simple model that used the assumption that the channel with the most favorable signal-to-noise ratio also determines behavioral performance. The current study evaluated this model psychophysically. First, several T/M spatial configurations were selected based on the criterion that they must have a narrowband spectral region with very favorable SNR (re. other spectral regions). The stimuli were then filtered so that they would activate mainly the peripheral channel with the most favorable SNR. Detection thresholds were then measured for the filtered and the unfiltered stimuli, both binaurally and monaurally. Large differences (up to 10 dB) in performance were observed, with binaural thresholds generally better than the corresponding monaural thresholds, which, in turn, were better than the single-channel thresholds. These results support the single-channel model only partially. However, they do not prove that across-channel integration plays a role in spatial release from masking.

1 Introduction

Detectability of a target sound (T) presented concurrently with another sound, a masker (M), is influenced, among other things, by the relative spatial position of the target and the masker. In most cases, spatial separation of T from M improves the target detectability, i.e., it leads to a spatial release from masking (SRM). Previous studies of SRM [1,2,3] suggested that, for non-speech targets masked by noise, two factors influence SRM: 1) the relative ratio of the target and the masker energy (TMR) in the peripheral filters of both ears, and 2) binaural processing (mostly for T stimuli with low-frequency content.

Lane et al. [1] performed a study of SRM in which they measured human performance when detecting a chirp-train stimulus masked by noise. They measured performance with broadband and lowpass-filtered stimuli, and tried to predict the data using a simple model (called the single-best-filter model, SBF) that only considered processing in a single peripheral channel - the one with the most favorable TMR, and ignored binaural, across-frequency or amplitude modulation processing. The SBF model accurately predicted broadband performance. However, the model was unable to predict the lowpass and the broadband data at the same time because the lowpass thresholds were worse than the broadband thresholds, while the model predicted identical performance. Lane et al. proposed that this discrepancy was due to across-frequency integration of the peripheral auditory information, which the model did not consider.

The goal of the present study was to more directly evaluate the hypothesis that across-channel integration is important in SRM, and that it was the missing integration part of the model that lead to the failure in the predictions of the Lane et al. data. We first replicated the results of the previous study, to make sure that possible differences in the results do not come from different experimental procedures. Then, we analyzed the outputs of the model peripheral filters for various spatial configurations of the T and M, and chose several prototypical spatial configurations. The selected configurations ranged from a configuration where one peripheral filter had clearly the most favorable TMR (and thus small effect of integration would be expected even if the integration was important) to a configuration where multiple channels had approximately equally favorable TMR (and thus there was plenty of opportunity for the across-channel integration to influence results). For the chosen spatial configurations the threshold TMRs were measured binaurally, monaurally, and with the target stimulus pre-filtered by the most-favorable model peripheral filter so that the across-channel integration is minimized. If across-channel integration improves performance then the pre-filtered thresholds were expected to be worse than the broadband thresholds. If not, then the thresholds were expected to be similar.

The present study was performed mostly with broadband stimuli for which the binaural contribution to SRM was expected to be small. This was important because otherwise, it might have been hard to distinguish the contribution of binaural processing

from the contribution of across-frequency integration, since both these factors were expected to improve performance.

2 Methods

2.1 Experimental procedure

The study consisted of two experiments. Five subjects with normal hearing participated in each experiment. Both experiments were performed in a virtual auditory environment, generated using non-individualized human head-related transfer functions (HRTFs). The target stimulus was a 200-ms long 40-Hz train of exponentially growing chirps with white spectrum in the range of 300-12,000 Hz (for lowpass conditions, the target was lowpass-filtered at 1,500 Hz). The masker was a white noise with frequency range of 200-14,000 Hz (for lowpass conditions, lowpass-filtered at 2,000 Hz). To determine the 79.8%-correct threshold TMR, 3-down-1-up 3-interval adaptive procedure was used, varying the T level. 3-interval, 2-alternative forced-choice procedure was used to collect responses. Stimuli were delivered via insert headphones in a quiet room.

In experiment 1 (described also in [5]), various azimuthal configurations of T and M were tested as indicated in Figure 1a. Most stimuli were broadband (except for three lowpass stimuli) and all were presented binaurally.

In experiment 2 (described also in [4]), only five azimuthal configurations were used (see panels "a" of Figures 2 to 6), chosen to examine the character of the across-frequency integration. For each spatial configuration, broadband binaural, broadband monaural, and several pre-filtered thresholds were measured (listed in panels "b" of Figs. 2 to 6). Gammatone filter [6] was used to pre-filter the signal so that the best-TMR peripheral auditory filter is activated most by the pre-filtered target.

2.2 Model

The "single-best-filter" model implemented to predict the data was identical to that used in the Lane et al. study [1]. The model computes the TMR in peripheral auditory channels a function of frequency, but does not allow for any across-frequency integration of information or any binaural processing. The model consists of a bank of 60 log-spaced gammatone filters [6] for each ear. For each spatial configuration, the root-mean-squared energy at the output of every filter is separately computed for the target and the masker. The model assumes that the filter with the largest TMR (over the set of 120) determines threshold. The only free parameter in the model, the TMR yielding 79.4% correct performance, was fit to match the measured threshold when broadband target and masker were at the same location.

3 Results

3.1 Experiment 1

The results of Experiment 1 are summarized in Figure 1. The main goal of this experiment was to compare the results obtained with the current experimental procedures to those of Lane et al.

Figure 1a shows the measured (symbols) and predicted (lines) thresholds as a function of the masker azimuth, for T azimuth fixed at 0°, 30°, or 90°. There is a very

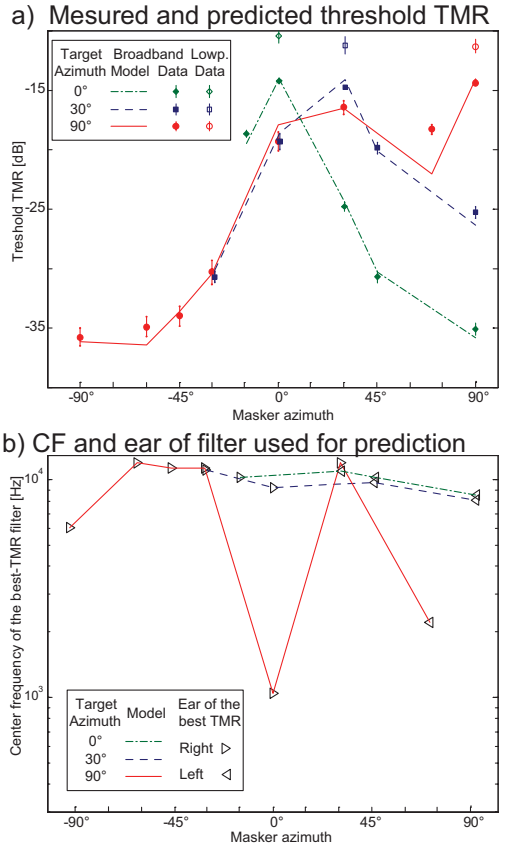

Figure 1: a) Measured and predicted threshold TMR for spatial configurations tested in Exp 1, plotted as a function of the Masker azimuth for a fixed target azimuth. b) Center frequency and ear (left vs. right) of the best-TMR filter based on which the corresponding prediction in panel a) was generated.

good match between the predicted and measured broadband thresholds. However, as in the previous study, the lowpass thresholds (open symbols) are consistently worse than the corresponding broadband thresholds. In addition, there is one broadband threshold (T @ 90°, M @ 70°, diamond symbol vs. the full line) mispredicted by the model. In this case, as well as in the lowpass cases, the model predicts that performance should be better than actually observed.

Figure 1b shows, for each broadband prediction from panel 1a, the ear (left or right) and the center frequency of the best-TMR peripheral filter on which the prediction was based. Most predictions were based on filters with high CF. However, the incorrect broadband prediction, as well as all the (incorrect) lowpass predictions, were based on filters with low CF. These results are very similar to results of Lane et al., suggesting that there might be a difference in the accuracy of predictions of high-CF vs. low-CF filters.

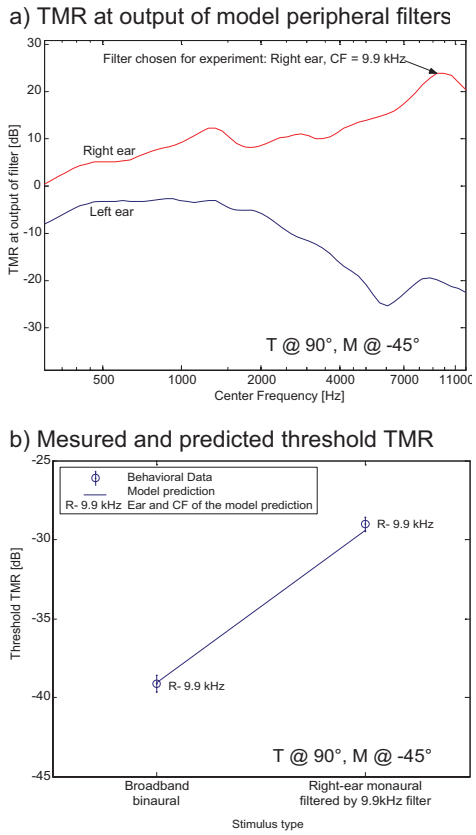

Figure 2: a) TMR in each model peripheral filter in both ears when T is at 90° and M at -45°. Arrow points to the filter that was chosen to for pre-filtering.
b) Measured (x-subject mean and SE) and predicted threshold TMRs for the stimulus conditions for which threshold was measured with T at 90° and M at -45°

3.2 Experiment 2

In experiment 2, thresholds were measured in 5 spatial configurations. Figure 2 describes the results obtained with T at 90° and M at -45°. This spatial configuration was chosen because there is a single high-CF peripheral filter for which the TMR is much better than for the other filters (see Figure 2a). To test whether this filter in deed determines performance, threshold was measured with broadband binaural target and with target and masker presented monaurally to the right ear, with the target pre-filtered by the chosen model filter.

Symbols in Figure 2b show the two measured thresholds. The broadband threshold is approximately 10 dB better than the threshold obtained with the monaural pre-filtered target. However, this difference is well described by the model (line), suggesting that the difference is not due to across-channel integration, but simply due to double filtering of the stimulus (first, the pre-filtering to generate the narrowband stimulus, and second, the actual peripheral auditory filtering).

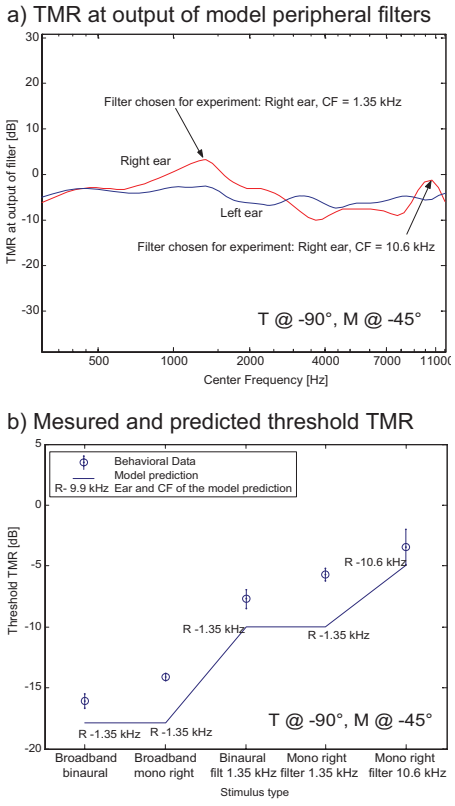

Figure 3: Description of figure identical to Figure 2. Spatial configuration with T at -90° and M at -45°.

The second chosen spatial configuration was T at -90° and M at -45°. As shown in Figure 3a, this configuration is interesting because there is a single low-CF dominant channel (there is also a relatively good high-CF channel that was included in the measurement to distinguish its potential contribution). Thus, across-channel integration, as well as binaural processing, might contribute to the broadband threshold.

Five different thresholds were measured in this spatial configuration (see Fig. 3b). The best threshold was obtained with broadband binaural presentation, followed by broadband monaural presentation. The narrowband pre-filtered thresholds were several dB worse than the broadband ones. The binaural thresholds are always a little bit better than the corresponding binaural thresholds, suggesting that there is some binaural contribution.

The model always predicted better performance than observed (lines vs. symbols in Fig. 3b). Since the broadband prediction is based on a filter with low CF, this model error is consistent with the errors observed in experiment 1. Moreover, if the model was fitted to the broadband thresholds (i.e., to the leftmost two data points), the narrowband thresholds would be predicted accurately (imagine shifting the whole line up by 3 dB). Thus, no across-channel integration is necessary to explain these data.

The third spatial configuration in Experiment 2 was T at -90° and M at 90°. As shown in Fig. 4a, in the left ear there are two high-CF channels with very favorable TMR. Three thresholds were measured (Fig. 4b), one broadband binaural, and one narrowband monaural for each of the two candidate channels. The results show that both the binaural and the 6-kHz monaural threshold are well predicted by the model, which means that considering the 6-kHz channel alone is sufficient to predict broadband performance. There is a large difference between the model prediction and the data for the 12-kHz threshold. This difference is probably a combination of inconsistent listener performance (note the large standard error bar) and some border effect, since the best filter is the filter with the highest CF considered.

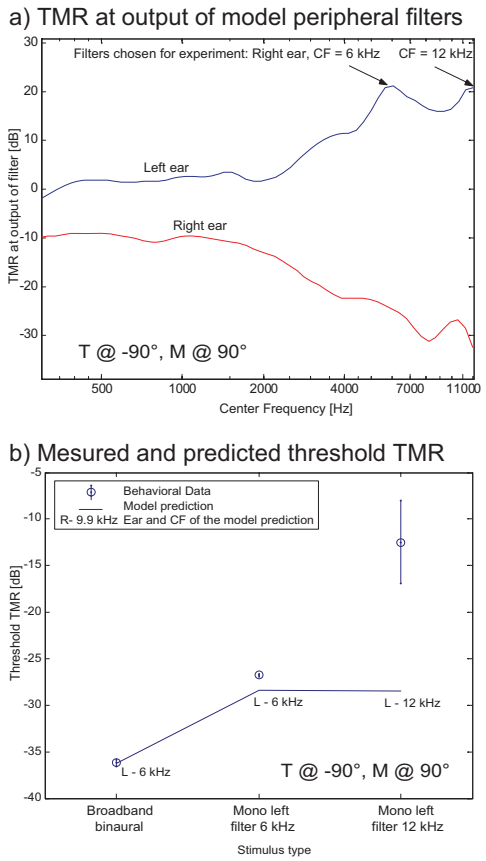

Figure 4: Description of figure identical to Figure 2. Spatial configuration with T at -90° and M at 90°.

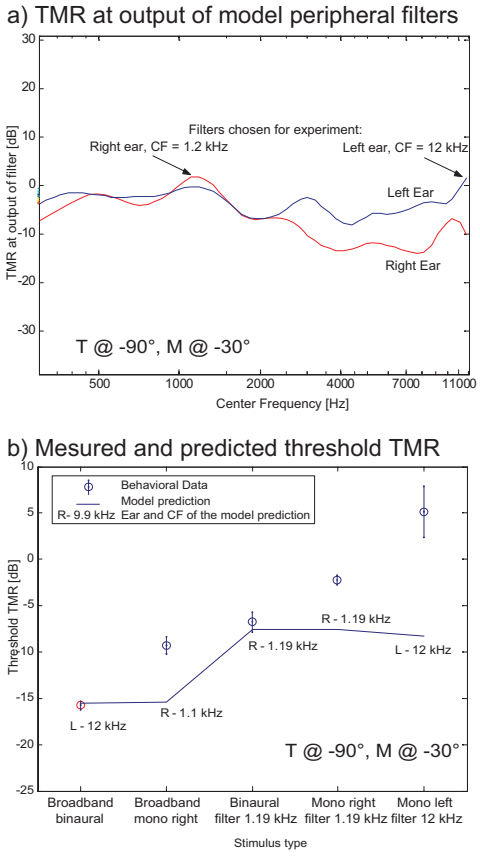

Figure 5: Description of figure identical to Figure 2. Spatial configuration with T at -90° and M at -30°.

The results obtained with T at -90° and M at -30° are shown in Figure 5. Fig. 5a shows that there are at least two channels with very good TMR in this configuration. The results show that binaural processing influenced the broadband binaural threshold (in Fig. 5b this threshold is much better than the others). However, there is still a good match between the two binaural thresholds and their predictions, as well as between the right-ear monaural thresholds and their predictions, suggesting that no across-channel integration needs to be evoked. Note that here again the left-ear monaural threshold is incorrectly predicted, probably for reasons similar to those discussed above for Figure 4.

The most challenging spatial configuration was that with T at 90° and M at 30° (Figure 6) where there are multiple low- and high-CF channels in both ears with approximately equal TMR (Fig. 6a). Figure 6b shows eight different measured and predicted thresholds,

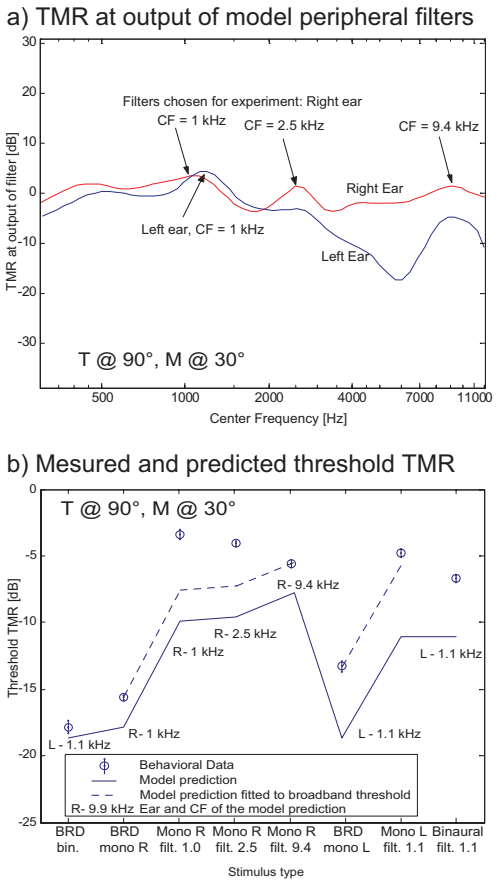

Figure 6: Description of figure identical to Figure 2. Spatial configuration with T at 90° and M at 30°.

considered in this configuration. First, comparison of the broadband binaural threshold (left-most circle) to the right-ear (second from left) and the left-ear (third from right) threshold shows that binaural processing contributed to the detection of broadband target. However, when binaural processing is accounted for by fitting the model to the monaural broadband thresholds (dashed lines), both left- and right-ear thresholds can be predicted by the best-TMR channels. Particularly interesting is the comparison of the predictions and data in the right ear (second through fifth point from the left in Fig 6b). Here, the broadband prediction is based on the low-CF channel, however, it is the high-CF channel that gives the lowest narrowband threshold. Moreover, the narrowband thresholds improve with increasing CF while the model predictions worsen with increasing CF, resulting in the low-CF threshold being much worse than predicted. This discrepancy is again consistent with the errors discussed above, in which the model had the tendency to predict better performance if the prediction was based on a low-CF channel.

4 Discussion and conclusions

The results of this study do not support the hypothesis that across-channel integration is necessary when considering spatial unmasking of stimuli with varying bandwidth. However, there were several occasions when contribution of binaural processing was observed, so considering binaural processing is important even for these broadband stimuli.

The only re-occurring error of the model was that the model tended to predict better performance than measured when the prediction was based on a low-CF channel. First, this error is probably not due to the model's lack of binaural processing or across-frequency integration, because both these mechanisms would make predictions even better, i.e., the error would be larger. Instead, the errors might be due to several other assumptions that the model makes. First, the model uses a gammatone filter bank to model peripheral processing. The observed errors in predictions might result from the gammatone filter being a more accurate model of auditory periphery at high frequencies than at low frequencies. Second, the model assumes that the threshold TMR is constant and independent of the filter CF. Again, assuming that the threshold TMR is higher at lower frequencies could correct the errors in predictions. And last, the stimuli used in this study produce 40-Hz amplitude modulation at the output of the peripheral filters. It might be that this modulation is used as a detection cue and that this cue is more efficient at higher filter CFs than at low CFs.

Further studies are needed to determine the actual source of this error, as well as to fully understand the

importance of modulation, binaural, and across-channel processing for spatial release from masking of non-speech and speech stimuli.

Acknowledgement

I would like to thank Branislav Beníkovský and Anton Baša whose diploma theses were important sources of information for this study. The research reported here was partially supported by grants from the Slovak Scientific Grant Agency (grant VEGA 1/1059/04) and the U.S. National Academy of Sciences.

References

[1] C.C. Lane, N. Kopco, B. Delgutte, B. G. Shinn-Cunningham, and H. S. Colburn. 'A cat's cocktail party: Psychophysical, neurophysiological, and computational studies of spatial release from masking' In: *Auditory signal processing: Physiology, psychoacoustics, and models.* (Pressnitzer, D., de Cheveigné, A, McAdams, S., and Collet, L., eds), pp 327-333, Springer, New York. (Proc. International Symposium on Hearing, Dourdan, France, Aug. 24-29, 2003)

[2] Saberi, K., Dostal, L., Sadralodabai,T., Bull, V., and Perrott, D.R. 'Free-field release from masking.' *J. Acoust. Soc. Am.* 90, 1355-1370. (1991)

[3] Good, M.D., Gilkey, R.H., and Ball, J.M. 'The relation between detection in noise and localization in noise in the free field.' In R.H. Gilkey and T.R. Anderson (Eds), *Binaural and Spatial Hearing in Real and Virtual Environments.* Lawrence Erlbaum Associates, Mahwah, N.J, pp 349–376. (1997)

[4] B. Beníkovský, '*Spatial unmasking of broadband stimuli in a virtual auditory environment*', Unpublished diploma thesis, TU Košice (2004)

[5] A. Baša, '*Význam priestorového vnímania a spracovania modulovaných podnetov pri počúvaní v komplexnom prostredí (Importance of spatial hearing and processing of modulated stimuli for listening in complex environments.)*', Unpublished diploma thesis, TU Košice (2005)

[6] Johannesma, P.I.M. 'The pre-response stimulus ensemble of neurons in the cochlear nucleus.' In: B.L. Cardozo, E. de Boer, and R. Plomp (Eds.), *IPO Symposium on Hearing Theory*. IPO, Eindhoven, The Netherlands, pp. 58-69. (1972).

Influences of modulation and spatial separation on detection of a masked broadband target[a)]

Norbert Kopčo[b)] and Barbara G. Shinn-Cunningham[c)]
Hearing Research Center, Boston University, Boston, Massachusetts 02215

(Received 22 June 2007; revised 24 June 2008; accepted 9 July 2008)

Experiments explored the influence of amplitude modulation and spatial separation on detectability of a broadband noise target masked by an independent broadband noise. Thresholds were measured for all combinations of six spatial configurations of target and masker and five modulation conditions. Masker level was either fixed (Experiment 1) or roved between intervals within a trial to reduce the utility of overall intensity as a cue (Experiment 2). After accounting for acoustic changes, thresholds depended on whether a target and a masker were colocated or spatially separated, but not on the exact spatial configuration. Moreover, spatial unmasking exceeded that predicted by better-ear acoustics only when modulation cues for detection were weak. Roving increased the colocated but not the spatially separated thresholds, resulting in an increase in spatial release from masking. Differences in both how performance changed over time and the influence of spatial separation support the idea that the cues underlying performance depend on the modulation characteristics of the target and masker. Analysis suggests that detection is based on overall intensity when target and masker modulation and spatial cues are the same, on spatial attributes when sources are separated and modulation provides no target glimpses, and on modulation discrimination in the remaining conditions. © *2008 Acoustical Society of America.* [DOI: 10.1121/1.2967891]

PACS number(s): 43.66.Dc, 43.66.Pn, 43.66.Rq, 43.66.Mk [RLF] Pages: 2236–2250

I. INTRODUCTION

The extent to which one sound source masks another depends to a large degree on how similar the two sources are in characteristics such as their spectral profile, temporal structure, and spatial location. While a fair amount is known about how these individual characteristics affect the ability to detect and understand a masked target, relatively little is known about how these characteristics interact. In everyday situations, listeners often are faced with the task of understanding one complex, fluctuating signal in the presence of similar, complex signals from different locations, such as understanding one talker in the presence of competing talkers. If we are ever to understand perception in everyday situations, we must explore how source characteristics such as spectral content, amplitude fluctuations over time (modulation), and spatial location jointly affect perception.

This paper considers the individual and combined effects of two stimulus characteristics: modulation structure and spatial location. *A priori*, one might imagine that the two variables are redundant with one other, so that there is no added benefit when spatial cues in a target and a masker differ if they already differ in their modulation structure (and vice versa). Alternatively, it is possible that masking effects related to temporal modulation and spatial location are largely independent of one another and that effects of the two attributes are additive. Finally, it is possible that differences in temporal modulations actually facilitate the effectiveness of spatial cues in releasing masking, or vice versa, resulting in superadditivity of their individual effects. This study investigates these alternative possibilities using a detection task with simple broadband noise targets and maskers by manipulating both temporal and spatial characteristics independently and jointly.

Several previous studies looked at spatial release from masking (SRM) for nonspeech stimuli that fluctuated over time. The target stimuli in these studies ranged widely, including click trains (Saberi *et al.*, 1991; Gilkey and Good, 1995; Good *et al.*, 1997), chirp trains (Lane *et al.*, 2004; Kopco, 2005), and pulsed 1/3-octave bands of noise (Zurek *et al.*, 2004). However, none of these studies looked at how modulation influences SRM.

Other studies examining the relationship between modulation and spatial processing in masked detection tasks differed substantially in approach and the specific questions addressed, making it difficult to compare results across studies. For example, some explored comodulation and binaural masking release (van de Par and Kohlrausch, 1998; Hall *et al.*, 2006) while others looked at monaural and interaural level discrimination (Stellmack *et al.*, 2005), the interaction between modulation detection interference and spatial processing (Sheft and Yost, 1997), or the equivalence of binaural processing of low-frequency fine time structure versus high-frequency envelope structure (Bernstein and Trahiotis, 1994; van de Par and Kohlrausch, 1997; Bernstein and Trahiotis, 2002). Physiological data from the cat inferior colliculus (IC) suggest that binaural cues in the temporal envelope contribute to SRM (Sterbing *et al.*, 2003; Lane and Delgutte, 2005). However, some psychophysical studies sug-

[a)]Portions of this work were presented at the 149th and 151st meetings of the Acoustical Society of America.
[b)]Permanent Address: Department of Cybernetics and AI, Technical University, Košice, Slovakia. Electronic mail: kopco@bu.edu
[c)]Electronic mail: shinn@bu.edu

gest that the stimulus temporal envelope does not affect SRM. For example, binaural detection thresholds obtained for a harmonic tone complex and broadband noise targets are very similar, despite dramatic differences in their envelopes (van de Par *et al.*, 2004). Overall, these studies do not provide a consistent account of how spatial cues and modulation jointly affect detection of a target embedded in noise.

Some work suggests that the influence of modulation on masked target detection depends on whether the target or the masker is modulated. For example, when listeners must detect a target embedded in maskers, reaction times depend less strongly on the number of distractors when the target is amplitude modulated and the maskers are unmodulated than when the target is a pure tone and the maskers are amplitude modulated (Asemi *et al.*, 2003). This asymmetry suggests that the modulated target is more likely to "pop out" of the background of unmodulated maskers than the reverse, making detection of a modulated target robust to the addition of interferers. In comodulation masking release (CMR) studies, adding off-target-frequency components that are modulated identically with the on-frequency masker improves the detectability of an unmodulated target (Hall *et al.*, 1984; van de Par and Kohlrausch, 1998; Winter *et al.*, 2004). However, we know of no studies reporting a corresponding benefit of increasing masker bandwidth when the target, rather than the masker, is modulated, so it is possible that there is a perceptual asymmetry between modulating the target versus modulating the masker in such situations, as well.

II. EXPERIMENTS AND HYPOTHESES

Two experiments were performed to study how modulation and spatial location of the target and masker affect target detection. Both target and masker were broadband noises that were either unmodulated or sinusoidally amplitude modulated (SAM). As a result, across-channel processing and across-frequency grouping were likely to contribute to performance. Moreover, for these broadband targets and maskers, listeners could not detect the target by using spectral sidebands (as might be the case when the target is a SAM tone; Dau and Ewert, 2004) and the opportunity to use profile analysis (Green, 1988) was minimized (because of the similarity of the target and masker spectral profiles).

A single modulation frequency (40 Hz) was used throughout the study, chosen both because humans are fairly sensitive to modulation at this frequency (Viemeister, 1979) and because responses of space-sensitive IC neurons are affected by modulation at this frequency (Lane and Delgutte, 2005).

Spatial separation of a broadband target from a broadband masker results in a frequency-dependent change in the target-to-masker energy ratio (TMR) at the ears. The resulting TMR profile as a function of frequency varies from one target/masker configuration to another, so that TMR should affect performance differently for different spatial configurations of the target and masker. The contribution of binaural processing to target detection should therefore depend on spatial configuration. In particular, if the TMR profile is such that the most favorable TMRs are at low frequencies, then interaural time difference (ITD) processing is likely to contribute to detection (Kopco and Shinn-Cunningham, 2003). On the other hand, if the most favorable TMRs are at high frequencies, then the contribution of ITD processing to performance is likely to be smaller. Finally, the contribution of across-frequency integration to detection, if any, is likely to be larger when the TMR is similar across frequency than when the TMR is very large in one band and small in others. As a result, the relative contribution of different detection cues (e.g., changes in overall energy and interaural decorrelation) also is likely to vary from one target/masker configuration to another.

Three different spatially separated configurations were included in this study to evaluate whether the interaction of modulation and spatial cues depends on the specific target/masker configuration. Specifically, in one of the chosen configurations the maximum in the TMR profile was in a low-frequency region, while in the remaining configurations it was at high frequencies.

As described above, the way in which modulation and spatial configuration interact is poorly understood. The current experiments were designed to explore how these cues jointly affect performance. If the processing of the two cues is strictly serial then the effects of the cues should be additive. This would occur if (1) spatial processing improves the effective TMR of the signal prior to any modulation processing, (2) modulation processing operates on the output of the spatial processing stage, and (3) detection is based on the output of the modulation processing. If the two cues both work to help listeners perceptually segregate the target from the masker, then the cues may be redundant. Specifically, if differences in modulation of the target and masker are sufficient to segregate the target and masker, then providing additional spatial cue differences in the target and masker might not improve performance. In this case, the benefits of modulation and spatial cue differences would be less than additive. Alternatively, if spatial cue differences are necessary for modulation differences to be useful (or vice versa), then the effects of differences in the two cues may be superadditive.

In addition to exploring whether the two cues are additive, subadditive, or superadditive, we tested two specific hypotheses about how source modulation structure and source location affect detection for broadband signals.

H1. The effect of modulation on SRM will depend on whether the target, the masker, or both target and masker are modulated (e.g., see the results of Asemi *et al.*, 2003).

H2. The effect of modulation on detection threshold will depend on spatial configuration because the relative importance of individual cues changes with spatial configuration. (1) When the best TMR occurs in low frequencies, ITD processing will be relatively influential on performance. (2) If perceived location rather than ITD processing is the critical factor in determining how spatial cues contribute to detection, performance will depend on whether or not the target and masker are spatially separated, but not on the exact spatial configuration. (3) When TMR is relatively constant with frequency, across-frequency integration is likely to contribute to detection.

Experiment 1 was performed with the masker noise presented at a fixed level. However, overall stimulus level may be the primary cue for detection when the target and masker are similar in their spectrotemporal structure and spatial cues, and therefore likely to be perceived as one unitary object from a particular location. To reduce the efficacy of overall level, Experiment 2 roved the masker level from interval to interval within each trial.

III. METHODS

A. Subjects

Seven subjects (four female and three male, including author N.K.) participated in Experiment 1. Seven subjects (three female and four male, two of whom participated in Experiment 1) participated in Experiment 2 (Experiment 2 was conducted almost a year after Experiment 1, so it is unlikely that learning from Experiment 1 transferred to Experiment 2 for the two subjects who performed both experiments). All subjects had normal hearing (confirmed by an audiometric screening), with ages ranging from 23–32 years.

B. Stimuli

The target and masker stimuli were both broadband noises with flat spectrum between either 0.3 and 8 kHz (target) or 0.2 and 12 kHz (masker), generated using a MATLAB implementation of the Butterworth bandpass filter (39th order for target and 33rd order for masker) with a stopband attenuation of 60 dB and stopband frequencies of 0.2–10.05 kHz (target) and 0.1–14 kHz (masker). The 200-ms-long target $s_T(t)$ was temporally centered on the masker $s_M(t)$, which had a duration of 300 ms. Both target and masker were ramped at onset and offset by 30 ms $\cos^2$ ramps. Modulation, if present, was sinusoidal with a frequency of 40 Hz and depth $m=0.5$ and had a random initial phase ϕ chosen from ten possible phases ($\phi=2\pi j/10$, $j=1,\ldots,10$). The stimuli were of the form

$$s_{i,k}(t) = A_i[1 + m_i \cos(2\pi 40t + \phi_{i,k})]n_{i,k}(t),$$

where $i=T$ for the target and $i=M$ for the masker, k is the trial number, $n_{i,k}(t)$ is a random bandpass-filtered noise token, and A_i is a scaling factor that determines the stimulus presentation level. The same five modulation conditions were explored in both experiments: no modulation ($m_T=m_M=0$), in-phase comodulation ($m_T=m_M=0.5$; $\phi_{M,k}=\phi_{T,k}$), target-only modulation ($m_T=0.5$; $m_M=0$), masker-only modulation ($m_T=0$; $m_M=0.5$), and pi-out-of-phase modulation ($m_T=m_M=0.5$; $\phi_{M,k}=\phi_{T,k}+\pi$).

Modulation increases the long-term rms energy of a signal by a factor of $(1+m^2)^{-0.5}$. For the modulation depth and form used here, modulation increases the rms energy of the modulated signal by approximately 0.5 dB. All results were corrected for this rms energy effect by scaling the measured thresholds and reporting thresholds in units of TMR.

Space was simulated using pseudoanechoic nonindividualized head-related impulse responses (HRIRs) recorded at four locations (−45°, 0°, 45°, and 90°, left to right) at a distance of 120 cm from the center of the head, using miniature microphones placed at the entrance of the ear canals of a female listener who did not participate as a subject in this study (see Shinn-Cunningham *et al.*, 2005, for a full description of these HRIRs). Five spatial configurations were explored in Experiment 1: two with the sources colocated at 0° or −45° and three with the sources spatially separated [(*T* at 90°, *M* at 0°), (*T* at 0°, *M* at 90°), and (*T* at 45°, *M* at −45°)]. An additional colocated condition (90°) was added in Experiment 2 to create three matching pairs of colocated and separated spatial configurations.

In both experiments, the average level of the masker was the same in all trials, prior to processing by the HRIRs (which altered the level of the signals reaching the ears). Therefore, because of HRIR processing, there were frequency-dependent variations in the signals reaching the ears across the different masker locations (graphs in Fig. 2 can be used to estimate how the received masker level changed at the two ears). For the masker at 0°, the maximum masker level received at the ears was 61 dB sound pressure level (SPL). In Experiment 1, the masker level was constant across the three intervals within a trial, while in Experiment 2 the masker level was roved independently in each interval by a value uniformly distributed between ±5 dB (the target, if present, was roved with the masker, which kept constant the TMR measured prior to HRIR processing).

Stimulus files, generated off-line at a sampling rate of 50 kHz, were stored on the hard disk of a control computer (IBM PC compatible). Ten random noise tokens were pregenerated to be used as targets and another ten tokens were produced to be used as maskers in this study (i.e., target and masker were always independent samples of noise). These 20 tokens were bandpass filtered (10 by the target filter and 10 by the masker filter, which had a slightly wider passband), modulated (by 1 of 10 modulation envelopes, differing in initial phase), and HRIR filtered (by an HRIR corresponding to locations of −45°, 0°, 45°, or 90°) to produce 440 target stimuli [10 tokens × (10 modulation envelopes + no modulation) × 4 locations] and 440 similar masker stimuli. On each trial, three different masker tokens and one target token were randomly selected, scaled, and concatenated into a stimulus file that contained three masker intervals with the target randomly added to the second or the third interval.

TDT System 3 hardware was used for D/A conversion. The result was amplified through a TDT headphone buffer and presented via Etymotic Research ER-1 insert earphones (with approximately flat frequency response in the range 100 Hz–15 kHz). No filtering was done to compensate for the transfer characteristics of the playback system. A simple alphanumeric interface in MATLAB was used to give instructions to subjects, gather responses, and provide feedback. The subject indicated the perceived target interval by hitting the appropriate numeric key ("2" or "3") on the computer keyboard. Experiments were performed in a single-walled sound-treated booth.

C. Experimental procedure

Each trial consisted of three intervals, each of which contained a masker. Either the second or the third interval

(randomly chosen with equal probability on each trial) also contained the target. The intervals were separated by 50-ms-long silent gaps. Subjects performed a two-alternative, forced-choice task in which they were asked to identify which interval, the second or the third, contained the target. Correct-answer feedback was provided at the end of each trial.

A three-down-one-up adaptive procedure was used to estimate detection thresholds (Levitt, 1971), defined as the 79.4% correct point on the psychometric function. Each run started with a description of the measurement condition of the run (e.g., written instructions might read "In this run, the target is modulated and the distractor is not modulated, the target comes from an azimuth of 0° and the distractor from 90°. Next, you will hear a sample of the noise distractor that you should ignore, followed by the target that you should identify. Hit RETURN to hear the sample."). The subject could listen to the sample repeatedly until he/she was confident that he/she understood the task.

The staircase measurement procedure started with the target presented at a clearly detectable level and continued until 11 "reversals" occurred. The target level was changed by 4 dB on the first reversal, 2 dB on the second reversal, and 1 dB on all subsequent reversals. For each adaptive run, detection threshold was estimated by taking the average target presentation level over the last six reversals.

Each of the two experiments consisted of six 1 h sessions performed on different days (the first session of each experiment was a practice session, serving to familiarize the subjects with the experimental procedure). In each session, the thresholds were measured for all combinations of spatial and modulation conditions (25 thresholds in Experiment 1 and 30 in Experiment 2), with the order of conditions randomized between sessions and between subjects. One adaptive run took approximately 2–3 min to complete.

Informal interviews of the listeners confirmed that at moderate to high TMRs, listeners found it very easy to interpret the two simulated stimuli as a target noise and a distractor noise coming from the indicated locations with the described modulation characteristics (as opposed to hearing them as one combined noise). This was likely the case because of the following: (1) at the beginning of the experiment, the subjects were given a detailed description of the stimulus combinations they should expect; (2) prior to each adaptive run, listeners had the opportunity to familiarize themselves with the target and masker stimuli presented separately before they heard them combined; and (3) the procedure started with both the target and the masker clearly audible. It is difficult to know whether or not the listeners perceived the two stimuli as separate objects when the target level was near the threshold. However, none of the subjects reported any difficulty performing the task (for example, none of them reported being confused about what to listen for in order to detect the target).

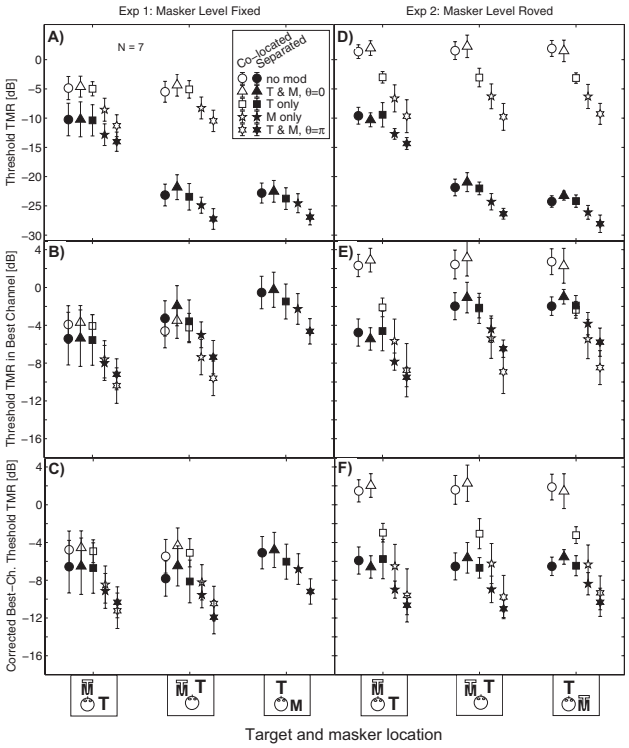

FIG. 1. Raw data plotted as a function of the masker location measured with the masker level fixed (Experiment 1; panels A, B, and C) and roved (Experiment 2; panels D, E, and F). All graphs show the across-subject mean and standard deviations in measured threshold TMRs: panels A and D show the raw threshold TMR energy ratios, panels B and E show the threshold TMRs in the best channel, and panels C and F show the threshold TMRs in the best channel after correcting for the frequency-dependence of the threshold TMR sensitivity.

IV. RESULTS AND ANALYSIS

A. Experiment 1: Fixed masker level

1. Overall results

Panels A, B, and C in Fig. 1 present the data collected in Experiment 1, with the masker level fixed (Panels D, E, and F show the data from Experiment 2, discussed in Sec. IV B). The data are plotted as a function of the masker location (indicated by the position of the letter "M" in the icons along the abscissa). Two spatial configurations are plotted for each masker location, one with the target and masker colocated (open symbols) and one with the target displaced from the masker (filled symbols).[1] The spatially separated target was at the location indicated by the filled letter "T" in the icons along the abscissa. The thresholds for different modulation conditions are represented by different symbols.

Figure 1(a) shows the across-subject mean and standard deviation of the TMR at detection threshold (lower values correspond to better performance). Thresholds varied by more than 20 dB, depending on the spatial configuration and modulation condition. For a given modulation condition and masker location, performance when the target and masker were spatially separated (filled symbols) was always better than when they were colocated (open symbols), revealing robust SRM. The colocated thresholds for target and masker at 0° and −45° were nearly identical, suggesting that the exact spatial configuration of the target and masker was not

important as long as the sources were colocated (this observation, based on the two configurations in Experiment 1, is further supported by the results of Experiment 2 in which three colocated thresholds were measured). In contrast, the spatially separated thresholds were strongly influenced by the specific target and masker locations: performance was worse with the masker at 0° than with the masker at −45° or 90° [compare the leftmost group of filled symbols in Fig. 1(a) to the center or the rightmost groups].

Within each spatial configuration, the no-modulation, in-phase comodulation, and target-only modulation (circles, triangles, and squares, respectively) thresholds were generally comparable, and these thresholds were higher (performance was worse) than the remaining thresholds. Masker-only modulation yielded improvements in performance (pentagrams fall below circles), while out-of-phase modulation of the target gave the lowest thresholds (hexagrams tend to fall below pentagrams).

A three-way repeated-measures analysis of variance (ANOVA) was performed with factors of modulation, spatial separation (colocated versus separated), and masker location (0, −45°), paralleling the layout of Fig. 1(a). The (*M* 90°, *T* 0°) configuration was omitted because it had no corresponding colocated measurement. This statistical analysis found a significant modulation × separation interaction ($F_{4,24}=7.63$, $p=0.0004$), a significant separation × masker location interaction ($F_{1,6}=950$, $p<0.0001$), and significant effects of all three main factors ($p<0.0001$). Notably though, neither the interaction between modulation and masker location nor the three-way interaction was significant ($p>0.1$). These results suggest that, although overall performance and the effect of separation depend on spatial configuration, at least for the spatial configurations explored in this study, the effect of modulation on the thresholds is similar within each spatial configuration rather than varying with target and masker locations.

2. Energy effects in 1/3-octave bands

One factor contributing to the large spatial benefits and to the dependence of these improvements on spatial configuration is the better-ear advantage, arising from the changes in the level at which the stimuli are received at the left and right ears when target and masker are spatially separated. In general, spatial separation of the target and masker sources produces a larger TMR at one of the ears (the "better ear"), and a smaller TMR at the other ear, compared to when the sources are colocated (where the TMR is equal at the two ears). To explore the extent to which changes in TMR at the acoustically better ear could account for the observed spatial unmasking, we calculated the TMR in each of the signals reaching the listeners' two ears as a function of frequency.

For each spatial configuration, we selected a target and a masker processed by the appropriate HRIRs and filtered both target and masker into 22 log-spaced 1/3-octave signals per ear (ANSI, 1986). In this analysis, the target and masker were set to have the same level prior to spatial processing. (Note that the effects of spatial processing on the TMR at the ears are identical for all modulation conditions.) The resulting frequency-dependent TMRs show the proper correction

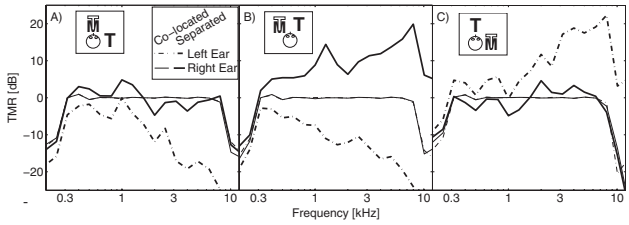

FIG. 2. TMR in 1/3-octave frequency bands in the six target/masker spatial configurations used in this study. Each panel shows the left- and right-ear TMRs in the colocated and separated spatial configurations for one masker location (indicated by the inset) as a function of the center frequency of the third-octave filters.

needed, at each frequency, to calculate the TMR at detection threshold in each of the 22 frequency bands.[2] The results of this analysis are plotted in Fig. 2.

Each panel in Fig. 2 shows the TMRs for one fixed masker location (indicated by the inset icon), with each combination of the ear (solid versus dashed lines for right versus left ear, respectively) and the spatial configuration (thin versus thick lines for colocated versus spatially separated) plotted separately. (Note that the dashed and solid thin lines lie nearly on top of each other, so only the solid thin lines are easily visible.)

TMRs for the colocated configurations (thin lines) were approximately zero (or less than zero at the edges where no target energy was present), independent of the masker location (across panels) or the ear (solid versus dashed thin lines). The spatially separated TMRs were frequency dependent and varied both with the ear (solid versus dashed thick lines within each panel) and with the masker location (panel A versus panel B versus panel C). The largest improvement in TMR with spatial separation was approximately 5 dB in panel A (right-ear channel centered at 1 kHz), approximately 20 dB in panel B (right-ear channel centered at 8 kHz), and approximately 22 dB in panel C (left-ear channel centered at 8 kHz). Assuming that the listeners detect the target by detecting its presence due to the energy effects in the frequency channel with the most favorable TMR, detection performance with spatial separation is expected to improve due to the spatial configuration by an amount equal to the maximum TMR shown in each panel of Fig. 2. Note that this analysis assumes that, in each condition, performance is determined solely by the single frequency channel with the most favorable TMR and that the threshold TMR calculated in 1/3-octave band is the same for all frequency channels. Therefore, this analysis ignores possible contributions of across-frequency integration and binaural processing. Moreover, the exact TMRs computed in this way will depend on the detailed shapes of the peripheral auditory filters used, as well as how they change with center frequency, so that slightly different corrections would be found with different filter assumptions. However, this analysis provides a first-order correction for the wide variation in TMR with frequency caused by HRIR processing.

Figure 1(b) shows the threshold TMRs in the best frequency channel, determined by adding the best-channel correction (i.e., the peak values from Fig. 2) to the respective thresholds in Fig. 1(a). Colocated thresholds [open symbols

in Fig. 1(b)] were essentially unchanged, as the TMR correction was near zero at all frequencies. However, correction of the spatially separated configurations reduced the effect of spatial separation to the point that many spatially separated thresholds (e.g., all thresholds with masker at −45°) were actually higher (performance was worse) than the corresponding colocated thresholds. Although this correction removed a good portion of the spatial effects on performance, ANOVA performed on the better-ear, best-frequency corrected thresholds found the same significant main factors and interactions as did the uncorrected thresholds [Fig. 1(a)] suggesting that the correction, while reducing the dependence of thresholds on the masker location, did not account for all of the variation in performance with spatial configurations.

3. Additional correction for frequency dependence of threshold TMR

The better-ear best-frequency correction yielded threshold TMRs that were much more similar than the uncorrected TMRs. To the extent that this correction was sufficient to account for the behavioral results, it suggests that (a) the threshold TMR is the same in all channels independent of frequency, (b) a simple 1/3-octave filter is an adequate representation of auditory filtering for the current analysis, and (c) there is no contribution of across-frequency integration or binaural processing to performance. The effect of any deviation from these assumptions is likely to depend on the spectral profiles of the target and masker signals, which differ with spatial configuration (see Fig. 2).

We now examine the assumption that threshold TMR in 1/3-octave band is constant as a function of frequency. In a previous study that measured SRM for broadband chirp-train signals masked by noise, threshold TMRs for narrowband targets were not constant as a function of frequency; instead, threshold TMRs were lower for higher-frequency targets (Kopco, 2005). When listening in a 9 kHz channel, best-channel analysis based on 1/3-octave filtering yielded thresholds that were nearly 4 dB lower than threshold TMRs using a 1 kHz channel. A simple frequency-dependent linear correction fit these earlier results relatively well (Kopco, 2005). The same correction, derived from the empirical fit to the data in this previous study, was applied to the current results:[3]

$$\mathrm{TMR}_{\mathrm{corrected}} = \mathrm{TMR}_{\mathrm{uncorrected}} + k_1 \mathrm{CF} + k_2. \quad (1)$$

Here, $\mathrm{TMR}_{\mathrm{uncorrected}}$ are the data from Fig. 1(b), CF is the center frequency of the best-TMR filter in Hz, the constant k_1 was fitted to Kopco's (2005) data (k_1 was estimated to be -4.9×10^{-4} dB/Hz), and the constant k_2 was arbitrarily set to 1.34 dB to minimize the offset of the corrected data from the raw colocated data. (Note that the constant k_2 does not influence relative comparisons, as it shifts all data points by the same amount, but simply accounts for the absolute value of the TMR threshold). The frequency-corrected best-TMR model uses the same assumptions as the best-channel TMR correction shown in Fig. 1(b), except that it relaxes the assumption of a constant frequency-independent threshold TMR sensitivity. Instead, threshold TMR is assumed to decrease linearly with increasing center frequency.

Figure 1(c) shows the thresholds corrected by Eq. (1). Compared to the graphs in Fig. 1(b), the corrected spatially separated thresholds [filled symbols in Fig. 1(c)] were always better than or equal to the corresponding colocated thresholds (open symbols). Thresholds were roughly equal across all masker locations [in Fig. 1(c), the *M* 0°, *T* 90° thresholds were approximately equal to the corresponding *M* −45°, *T* 45° thresholds, as well as to the *M* 90°, *T* 0° thresholds; the trend was confirmed by data shown in Fig. 1(f) from Experiment 2]. Because the same correction was applied to all thresholds for a given spatial configuration, independent of the modulation condition, colocated thresholds still changed more as a function of the modulation condition than did the spatially separated thresholds. (Supporting these observations, ANOVA performed on the corrected data only found one significant interaction, modulation × separation, $F_{4,24}=7.65$, $p<0.0005$; all three main effects were significant, with $p<0.05$.) With these corrections, the spatially separated thresholds were only consistently lower than colocated thresholds in the no-modulation, in-phase modulation, and target-only modulation conditions [filled versus open circles, triangles, and squares in Fig. 1(c)]. Colocated and spatially separated thresholds were statistically indistinguishable in the masker-only modulation and out-of-phase modulation conditions for all spatial configurations.

Given the similarity of the corrected best-channel threshold TMRs at different masker locations [Fig. 1(c)], there only appears to be a modest effect of across-frequency integration in this study (i.e., there are no large differences across different spatial configurations, even though the best frequency and the overall shape of the better-ear TMR as a function of frequency vary dramatically with spatial configuration). Similarly, spatial processing only appears to contribute when the masker is modulated in a way that does not provide glimpses of the target (in the no modulation, in-phase modulation, and target-only modulation conditions).

In all of the following sections, the frequency-corrected best-channel TMR thresholds [from Figs. 1(c) and 1(f)] are used because (1) this correction accounts for the dependence of the thresholds on the masker location; (2) even though consideration of binaural processing and across-frequency integration could also produce corrections that explain some of the variability as a function of the masker location,[4] parsimony argues that these factors played only minor roles in this experiment; and (3) the fact that spatially separated configurations produce thresholds that depend less on the modulation condition than do colocated configurations is independent of the method used to account for energy effects or of the masker location. (However, note that it is currently not clear what causes the frequency dependence of the 1/3-octave filtered threshold TMRs.)

4. Results collapsed across the masker location

To better assess the interaction between modulation and separation, Fig. 3 shows the data collapsed across masker location. Figure 3(a) plots the across-subject mean threshold TMRs in the best 1/3-octave channel (and within-subject standard deviation, chosen here because it removes the between-subject differences from the computation of stan-

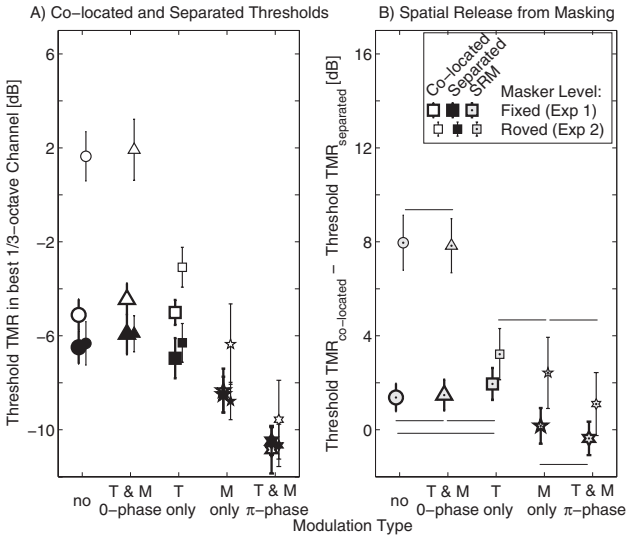

FIG. 3. Threshold TMRs in the frequency-corrected best 1/3-octave channel (panel A) and SRM (panel B) as a function of the modulation type, averaged across the masker locations (error bars give the within-subject standard deviation). The horizontal lines in panel B indicate SRMs that were not significantly different at the 0.01 level in a t-test after correcting for multiple comparisons (lines below large symbols for Experiment 1; lines above small symbols for Experiment 2). Different symbols are used to identify the modulation type, as in Fig. 1. The legend in panel B applies to both panels and all modulation conditions.

dard deviation[5]) as a function of the modulation type. The large filled and open symbols represent the spatially separated and colocated thresholds, respectively (the small symbols represent the results of Experiment 2, discussed in Sec. IV B).

The effect of modulation on performance was similar for colocated and separated spatial configurations. Thresholds were essentially the same for the no-modulation, in-phase comodulation, and target-only modulation conditions [compare large open and filled circles, triangles, and squares in Fig. 3(a)]. Performance with masker-only modulation (pentagrams) and out-of-phase modulation (hexagrams) was better, with lower thresholds.

Although the rank ordering of thresholds was the same for colocated and spatially separated conditions, the dependence of the thresholds on modulation was slightly stronger when the sources were colocated than when they were spatially separated (large open symbols span a range of nearly 7 dB, while the large filled symbols span a range of about 4 dB), suggesting that spatial separation affects performance differently for different modulation conditions. This SRM [the difference between the open and filled symbols in Fig. 3(a)] is plotted as a function of the modulation condition in Fig. 3(b). This panel shows the across-subject mean (and the within-subject standard deviation[5]) of the difference between the spatially separated and corresponding colocated thresholds from panel A.

One-way repeated-measures ANOVA found a significant effect of modulation on SRM ($F_{4,24}=11.44$, $p<0.0001$). The results of Bonferroni-corrected *post hoc* pairwise t-tests (which account for heterogeneity of variances; e.g., Ury and Wiggins, 1971) as implemented in the CLEAVE package (Hetron, 2005) are also shown in Fig. 3(b). The horizontal lines under the large symbols in Fig. 3(b) indicate those pairs of conditions in Experiment 1 that did not differ at the $p<0.01$ significance level (all other pairs were significantly different from one another). The no-modulation, in-phase modulation, and target-only modulation SRMs were not significantly different from one another. Similarly, the masker-only modulation versus out-of-phase modulation SRMs were not significantly different from one another. However, the modulation type had a small but significant effect on the SRM: compared to no-modulation, in-phase modulation, or target-only modulation [circle, triangle, and square in Fig. 3(b)], modulating only the masker (pentagram) or modulating the target and masker stimuli with opposite phases (hexagram) decreased the SRM by roughly 1.5–2 dB ($p<0.01$), resulting in no benefit of spatial separation in the latter modulation conditions.

Finally, as discussed in the Appendix, learning affected SRM: in the first of the five repeats of this experiment, the SRM was essentially the same for all types of modulation (the largest difference was less than 1 dB). However, by the fifth repeat, the difference between the target-only modulation and the out-of-phase modulation grew to more than 4 dB. Thus, the average effect plotted in the data collapsed across the repeats is smaller than might be seen after extensive training.

B. Experiment 2: Masker level roved

To isolate the contribution of the overall level cue to performance, Experiment 2 was performed with the masker level roved between the intervals within a trial, a strategy used extensively in the profile analysis literature (Mason *et al.*, 1984; Kidd *et al.*, 1989). The (T 90°, M 90°) colocated condition was added to balance the number of colocated and spatially separated conditions; otherwise, Experiment 2 was identical to Experiment 1, except with a random ±5 dB intensity rove added from interval to interval.

1. Overall results

Panels D, E, and F in Fig. 1 present the results of Experiment 2 in a format identical to Experiment 1 (see Sec. IV A). The raw data in Fig. 1(d) followed the same trends as in Experiment 1. The spatially separated thresholds (filled symbols) were almost identical to those found in Experiment 1. The colocated thresholds for the no-modulation (circles) and in-phase modulation (triangles) conditions tended to be worse than in Experiment 1. However, the level rove had little effect on the remaining colocated configurations (a direct comparison is presented below). This result suggests that overall level was the main cue used for detection only in the colocated configurations in which the target and masker had identical temporal envelopes, a conclusion that was confirmed by a comparison of the data in panels E and F to respective panels B and C. (ANOVAs performed on the raw and corrected Experiment 2 data from panels D, E, and F found the same significant main effects and interactions as the respective ANOVAs performed on the Experiment 1 data.)

2. Results collapsed across the masker location

In order to analyze the interaction between modulation and spatial separation, the data were collapsed across the masker locations. To allow a direct comparison of the effect of masker level uncertainty, Fig. 3 shows the results for Experiment 2 (the small symbols slightly offset to the right) plotted alongside the data from Experiment 1 (larger symbols).

The filled symbols in Fig. 3(a) show the spatially separated thresholds. Roving the masker level had essentially no effect on any of the spatially separated thresholds (compare the small and large filled symbols from Experiments 2 and 1, respectively). In contrast, all colocated thresholds were larger in Experiment 2 than in Experiment 1 (the small open symbols fell above the corresponding large open symbols). The largest increase (around 7 dB) was observed when the target and masker had identical temporal envelopes (i.e., in the no-modulation and in-phase comodulation conditions; circles and triangles). In the three remaining modulation conditions, the masker-level rove increased thresholds by approximately 2 dB.

Figure 3(b) shows that, as a consequence of the effects of the level rove on the colocated configurations, the SRM was much larger in Experiment 2 than in Experiment 1 in the conditions in which the target and masker had the same temporal envelope. A one-way repeated-measure ANOVA revealed a significant effect of modulation on the SRM ($F_{4,24}=35.55$, $p<0.0001$). Bonferroni-corrected *post hoc* pairwise t-tests found no significant differences between unmodulated and comodulated SRMs, target-only and masker-only modulated SRMs, or the masker-only and out-of-phase modulated SRMs (see the horizontal bars above the pairs of small symbols that were not significantly different; $p>0.01$). All other pairs of modulation conditions showed statistically significant differences.

The results in Fig. 3 suggest that overall level was used to detect the target when the colocated target and masker had the same envelope. For wideband noise, the smallest detectable intensity change ΔI is proportional to the base line intensity, I, so that $\Delta I/I$ is approximately constant with values between −9 and −6 dB over a large range of I (20–100 dB above the absolute thresholds; Moore, 2003). The results for colocated identically modulated stimuli in Experiment 1 match these data well, with TMR thresholds of approximately −5 dB [large open circles and triangles in Fig. 3(a)]. If overall level was the only available cue in this two-alternative forced-choice task and the external noise of the 10 dB rove dominated performance, then the TMR at detection threshold would be 1.07 dB for an ideal observer (Durlach *et al.*, 1986; Green, 1988), which is remarkably close to the actual thresholds observed for the identically modulated and in-phase modulated conditions, where threshold TMRs were around 2 dB. In most previous studies of the effect of rove on profile analysis, the rove yielded performance that was worse than was predicted for an optimal observer (Spiegel *et al.*, 1981; Mason *et al.*, 1984). Thus, even the fact that thresholds are slightly higher than the ideal-observer prediction is consistent with past work. Moreover, the no-modulation and in-phase comodulation thresholds were very similar to each other, suggesting that the fluctuating envelope in the latter condition did not make it harder to judge the levels in the different intervals.

In conditions for which target and masker were colocated but had different temporal envelopes, performance was much better than would be predicted if the main cue used for target detection was overall intensity, showing that some other nonlevel cue was the main feature used to detect the target. Nevertheless, in such conditions, the rove interfered slightly with performance, a result that suggests that the intensity rove made it more difficult for listeners to extract whatever feature was the main detection cue when target and masker were colocated.

C. Modulation detection

To understand the effects of modulation on performance, two analyses were performed. First, the instantaneous TMR was analyzed. In this analysis, predictions were based on detecting the target by hearing its effect at the best instant in time. A second analysis assumed that the listeners detected the target+masker interval by detecting a modulation depth that was different from the masker-only modulation (in the nontarget intervals).

1. Listening at peaks and dips: Instantaneous TMR analysis

The presence of modulation in the stimuli caused the instantaneous TMR to change over time. Humans appear to utilize these changes and detect the target in moments when the TMR is most favorable, both in monaural (Buus *et al.*, 1996) and binaural (Buss *et al.*, 2003) listening tasks, even though this ability can differ across subjects (e.g., see Buss *et al.*, 2007). Of course, given that the ability to utilize these cues is limited by the temporal resolution of the auditory system, factors like forward masking are likely to influence the ability to listen in dips (Widin *et al.*, 1986; Wojtczak and Viemeister, 2005). While the present analysis does not consider these limitations, it does provide an upper limit on how much the listeners could have benefited from changes in the instantaneous TMR. Specifically, if one assumes that the peak TMR produced after temporal integration over some fixed time window predicts performance, the current analysis gives the limit of performance if temporal resolution is infinitely precise, leading to an effective time window that is infinitely narrow. Conversely, the overall-TMR analysis shown in Fig. 3(a) shows predictions for an infinitely long time window. Any finite-length time window must produce results intermediate between these two extremes.

In the colocated conditions with identical modulation (no modulation and in-phase comodulation; circles and triangles), the TMR was constant over the duration of the stimulus. In the conditions with different target and masker modulations, the difference between the long-term TMR and the peak instantaneous TMR depended on which stimulus was modulated. Because the modulation envelope was sinusoidal in pressure units, the effect of modulation on the instantaneous sound pressure level was not symmetrical in decibel units. For sinusoidal modulation with a modulation

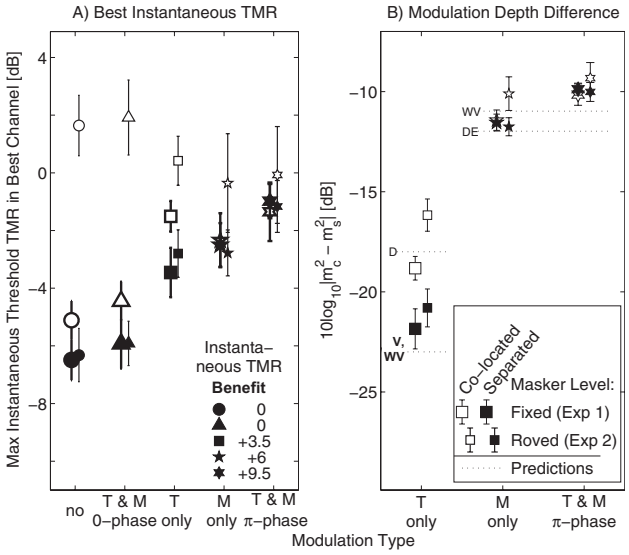

FIG. 4. (A) Peak instantaneous TMRs at threshold in the best 1/3-octave channel [derived from Fig. 3(a) by applying the instantaneous TMR benefit corrections, listed in the inset, to both colocated and separated thresholds of both Experiments 1 and 2] (B) Modulation depth (across-subject means and within-subject 95% confidence interval) at the threshold TMR in the three modulation conditions in which modulation of the target and masker differed. Data are compared to predictions based on the data of Wakefield and Viemeister (1990)—WV, Dau and Ewert (2004)—DE, Viemeister (1979), and Dau (1996)—D. The legend of panel B applies to data in both panels and to all modulation conditions.

depth of 0.5 (used in this study), the instantaneous signal level at the minima of the modulation envelope was 6 dB lower than the level with no modulation, while the level at the peaks of the modulation envelope was 3.5 dB higher than the unmodulated level.

Figure 4(a) plots the best instantaneous TMR in the frequency-corrected best 1/3-octave channel at threshold, determined by adding the instantaneous-TMR-benefit corrections (described above and listed in the inset) to the long-term frequency-corrected TMR thresholds in the best frequency channel [from Fig. 3(a)]. [Note that for each modulation condition, colocated and spatially separated thresholds have the same instantaneous-TMR-benefit correction, so that this correction does not influence SRM, shown in Fig. 3(b).]

As seen in Fig. 4(a), the peak instantaneous TMR at detection threshold falls between −4 and 0 dB for the conditions in which the target and masker envelopes differ [target modulation, masker modulation, and out-of-phase modulation conditions; large open squares, pentagrams, and hexagrams in Fig. 4(a)]. These values are higher than the intensity just noticeable difference (JND) (−9 to −6 dB, as discussed above), suggesting that listeners were unable to make use of the peak instantaneous TMR to detect the target based on changes in overall intensity. Given that the long-term average TMR does not capture the differences in thresholds as a function of modulation type [if it did then the thresholds represented by the large open squares, pentagrams, and hexagrams would be constant in Fig. 3(a)], while the instantaneous TMR predicts performance that is too poor (even though it is approximately constant), it is possible that predictions based on the TMR averaged over an appropriate finite-length time window could account for detection based on changes in intensity. However, if performance were based on the same intensity cue for cases when target and masker had the same envelope and cases when the target and masker envelopes differed, the effect of intensity rove should be similar in all conditions. Instead, intensity rove affected performance in the different conditions very differently, suggesting that some cue other than overall intensity integrated over some finite-duration time window enabled target detection when target and masker envelopes differed.

2. Effect of the target on the masker envelope modulation

One attribute that is affected by the addition of the target to the masker is the shape of the total stimulus envelope (Dau *et al.*, 1997). The salience of any change in the envelope due to the presence of the target depends on the relative levels of the target and masker as well as on the modulation condition. In the target-only-modulated condition, modulation is only present in the target interval and listeners may detect the target by detecting the presence of modulation. In the masker-only-modulated and the target-and-masker-modulated-out-of-phase conditions, the addition of the target decreases modulation depth from the 0.5 depth in the nontarget intervals and listeners may discriminate changes in the modulation depth to detect the target.

Detection and discrimination thresholds for modulation can be expressed as the modulation index $10\log_{10}|m_c^2-m_s^2|$, where m_s represents the modulation depth of the standard (i.e., in the nontarget interval) and m_c is the modulation depth of the stimulus at discrimination threshold (i.e., the modulation depth of the combined target+masker signal in the target interval). The current target-modulated thresholds can be estimated either from previous modulation detection data (Viemeister, 1979; Dau, 1996) or from discrimination data using a standard with a very low modulation depth (Wakefield and Viemeister, 1990, and Dau and Ewert, 2004; summarized in Fig. 2 of Dau and Ewert, 2004).[6] For modulation detection, the modulation index at threshold is in the range from −23 dB (Viemeister, 1979) to −18 dB at threshold (Dau, 1996). The results from modulation discrimination experiments (Wakefield and Viemeister, 1990) suggest that modulation index thresholds are near −23 dB for standard depths less than −30 dB.

Thresholds from a previous modulation discrimination study (e.g., Dau and Ewert, 2004) can be linearly approximated as $10\log_{10}(m_c^2-m_s^2)=10\log_{10}m_s^2-4$, from which the predicted threshold for a decrease in modulation from the standard of $m_c=-6$ dB can be estimated as $10\log_{10}(m_c^2-m_s^2)=-11$ dB [thresholds from Wakefield and Viemeister (1990) are approximately 1 dB larger than the Dau and Ewert (2004) thresholds when analyzed in this way].

In order to compare the current data to these predictions, the relationship between the threshold TMRs and the modulation depth of the combined stimulus was examined for our stimuli. However, combining a SAM noise and an unmodulated noise does not produce a stimulus with sinusoidal amplitude modulation. The relation between the threshold

modulation and threshold TMR was estimated by determining the maximum and minimum amplitudes, of the combined stimulus envelope and then finding the modulation depth of a SAM stimulus that would give the same maximum and minimum (although the exact shape of the modulation envelope differs, the difference is relatively small, especially near threshold). The resulting relationships for the three differential modulation conditions in this study (and for the target and/or masker modulation of 0.5) are as follows.

In target-only modulated,

$$m = \frac{\sqrt{1 + 1.5^2\mathrm{TMR}^2} - \sqrt{1 + 0.5^2\mathrm{TMR}^2}}{\sqrt{1 + 1.5^2\mathrm{TMR}^2} + \sqrt{1 + 0.5^2\mathrm{TMR}^2}}.$$

In masker-only modulated,

$$m = \frac{\sqrt{1.5^2 + \mathrm{TMR}^2} - \sqrt{0.5^2 + \mathrm{TMR}^2}}{\sqrt{1.5^2 + \mathrm{TMR}^2} + \sqrt{0.5^2 + \mathrm{TMR}^2}}.$$

In stimuli modulated out of phase,

$$m = \frac{\sqrt{1.5^2 + 0.5^2\mathrm{TMR}^2} - \sqrt{0.5^2 + 1.5^2\mathrm{TMR}^2}}{\sqrt{1.5^2 + 0.5^2\mathrm{TMR}^2} + \sqrt{0.5^2 + 1.5^2\mathrm{TMR}^2}},$$

where TMR is the threshold TMR in the best channel (from Fig. 3) in pressure units and m is the threshold modulation depth of an equivalent SAM noise. These equations can be inverted to estimate the target+masker modulation depth at target detection threshold for the measured results.

Figure 4(b) shows data for the three modulation conditions in which target modulation is different from the masker modulation, expressed as the difference in modulation depth between the target+masker interval and the reference masker-alone interval (the modulation conditions for which the target and masker have the same envelope were not included in this analysis because there is no change in modulation with addition of the target). Also shown are the predictions estimated from results of Viemeister (1979), Wakefield and Viemeister (1990), Dau and Ewert (2004), and Dau (1996; see dashed lines).

The thresholds for the colocated stimuli with fixed masker levels (open large symbols) generally match the previous detection and discrimination data fairly well for all three types of modulation, suggesting that the listeners detected changes in modulation depth in these conditions. The spatially separated thresholds are only lower (detection is easier) than the colocated thresholds in the target-modulation condition, when the listeners do not ever get a good "glimpse" of the target (large filled versus open squares). At first glance, the fact that the spatially separated thresholds fall within the range of the previous modulation detection data (i.e., between the dotted lines marked by D and V, VW) seems to suggest that the listeners did not benefit from spatial cues in this condition. However, given the large difference between the D and the V, VW thresholds, and given that there is a consistent difference between the colocated and spatially separated thresholds in the current study, it is clear that the listeners did use the spatial separation cue, in addition to modulation, here.

Finally, although the effect is small, colocated roved thresholds (open small symbols) consistently fall above the range of thresholds observed in previous studies which did not rove the stimulus presentation level. This shows that overall level rove impaired the listeners' ability to detect or discriminate modulation in the current study.

V. DISCUSSION

Noise-on-noise threshold TMRs changed over a range of 30 dB [Figs. 1(a) and 1(d)], and were influenced by the spatial configuration of the target and masker, the type of modulation present in the stimuli, and a rove of the masker level. Moreover, as discussed in the Appendix, these differences appear to increase with experience. A large part of the variability in performance across the tested conditions (as much as 20 dB) came from the changes in the target and masker energy levels received at the ears when the target and masker locations changed. Specifically, if one considers the TMR within the best 1/3-octave frequency channel in the acoustically better ear, threshold TMRs ranged only over 5 dB across different spatial configurations. If one then corrects these detection thresholds based on the detection threshold differences across frequency,[3] threshold TMRs were even closer, spanning a range of only about 1 dB across the different spatial configurations for a given modulation condition.

As shown in Fig. 2, the way in which TMR varies with center frequency differs dramatically across the spatial configurations used in this study. Therefore, any contributions of ITD and across-frequency processing to performance are likely to depend on masker location. However, no large differences were observed after applying frequency-dependent corrections to the TMR in the best frequency channel. Thus, for the broadband stimuli used here, both binaural and across-frequency contributions to performance appear to be modest. Frequency-dependent TMR thresholds could also explain the results of a previous related experiment without considering any across-frequency integration or binaural processing (Lane *et al.*, 2004). Together, these results suggest that low-level binaural processing does not contribute very much to spatial unmasking when detecting a broadband target in a broadband masker (although it can contribute significantly when the target is narrowband; e.g., see Kopco and Shinn-Cunningham, 2003).

The benefit of spatial separation found in the current results is similar for all spatial configurations, even though the best frequency channel is sometimes in a low-frequency region where binaural processing is expected to provide a large benefit and sometimes in a high-frequency region where binaural processing typically provides much more modest benefits (Zurek, 1993; Kopco and Shinn-Cunningham, 2003). This suggests that differences in the perceived *spatial attributes* of the stimuli (which depend both on low-frequency ITDs as well as high-frequency interaural level differences and spectral cues) are responsible for the spatial unmasking not explained by changes in the TMR at the better ear, rather than *binaural processing* that operates primarily at low-frequencies (unmasking caused by interau-

ral decorrelation; Colburn, 1977); (see Freyman *et al.*, 1999, for another study contrasting how spatial perception and binaural processing contribute to spatial unmasking).

Both modulation and intensity rove influenced the SRM, defined as the difference between the best-channel threshold TMRs with colocated and spatially separated stimuli. With the masker level fixed, SRM was comparable for no-modulation, target and masker in-phase modulation, and target-only modulation configurations, but SRM was statistically insignificant when only the masker was modulated or target and masker were modulated out of phase [see Fig. 3(b)]. Uncertainty about the masker level increased SRM in all modulation conditions, but the size of this effect depended on the modulation in the stimuli. For the level-roved stimuli, SRM was 7 dB larger when the target and masker have the same temporal envelope, but only 2 dB larger when the stimuli had different modulation. These results can be understood by considering how and when listeners use overall level, modulation, and spatial cues to detect the presence of the target.

A. Overall level

Detection in the colocated, identically modulated conditions [i.e., when neither modulation nor spatial cues were available for target detection; open circles and triangles in Fig. 3(a)] appears to be based on detecting changes in overall intensity. This conclusion is supported by (1) the observed good match between thresholds in these conditions and predictions from previous intensity JND studies (Experiment 1) and (2) the effect of the intensity rove in these conditions (Experiment 2), which increased detection thresholds to just above that expected for an ideal observer using overall level as the detection cue (Green, 1988). (However, note that there were small gating asynchronies and spectral differences between the target and masker signals that could have contributed to the detection of colocated identically modulated targets.)

B. Space cue alone

When stimuli differed in their spatial locations but not in their modulation [filled circles and triangles in Fig. 3(a)], a consistent improvement in performance was observed, showing that spatial separation provided benefits beyond the improvements in the better-ear TMRs. Changes in the spatial attributes of the target+masker versus masker-only stimuli (such as perceived spatial width) likely were used to detect the target at threshold, a conclusion particularly supported by the fact that the threshold was not influenced by the intensity rove [large and small filled circles and triangles are the same in Fig. 3(a)].

C. Modulation cue alone

Differences in the target and masker modulations led to some improvements in detection when the target and masker had the same location, but not in all conditions. Modulation led to lower thresholds when only the masker was modulated and when the target and masker were modulated out of phase, independent of whether the overall level was roved or not [compare open pentagrams and hexagrams to open circles and triangles in Fig. 3(a)]. When the level was roved, modulation also improved detection when only the target was modulated [compare small open square to small open circle and triangle in Fig. 3(a)]. However, when the level was fixed, the target-only modulation did not improve performance compared to when there were no modulation cues to detect the target [compare large open square to large open circle and triangle in Fig. 3(a)].

The intensity rove caused modest degradations in performance when colocated target and masker had different modulation envelopes, hinting that the listeners might have used the overall level cue (selected at the most favorable TMR instances) instead of the modulation cue in these conditions. However, given that the rove effects were much smaller than when target and masker had identical envelopes, and that the thresholds in these cases were better than (i.e., below) those predicted for an ideal observer using intensity increments to detect the target (Green, 1988), it is unlikely that the listeners used overall level to detect the presence of the target in these conditions [small open squares, pentagrams, and hexagrams in Fig. 4(a)]. Instead, it seems that roving overall level made it slightly harder to judge the changes in modulation caused by adding a target to a masker in these tasks. However, in the target-only modulation condition, the long-term TMR threshold is comparable to that for the no-modulation and in-phase modulation conditions when the level is fixed [large open square, triangle, and circle are comparable in Fig. 3(a)]. Moreover, when the level was not roved, the spatial separation improved performance by similar amounts when only the target was modulated and in the cases where the level was clearly the cue for detection (no modulation, in-phase modulation). Thus, for the target-only modulation condition, it is possible that the subjects used an overall level to detect the target when the level was roved and used a modulation to detect the target when the level varied randomly from interval to interval.

Another result hinting that the subjects' behavior might have been more complex than just detecting the modulation depth is that no similar effect of an intensity rove was seen in a previous study that measured modulation discrimination (Stellmack *et al.*, 2006). However, this difference in the effect of an intensity rove in the two studies may be due to the differences in the instructions given to subjects. In the previous study, listeners were instructed to detect changes in the modulation depth of a single stimulus, while in the current study they were presented with examples of the masker and target at the start of each block and instructed to detect the presence of the target. This priming may have enhanced the likelihood that listeners perceptually segregated the target from the masker in the current study, or that they switched cues between the rove and no-rove experiments, rather than detecting the target+masker interval by perceiving a change in masker attributes. However, further experiments are required to explore which of these alternatives is correct.

D. Space and modulation

Spatial separation did not always improve detection beyond performance for colocated sources after accounting for the TMR at the best frequency in the better acoustic ear. Specifically, spatial separation did not improve detection very much, other than by changing TMR, when the masker envelope had dips, providing good glimpses of the target (in the out-of-phase and masker-only modulation conditions). As noted above, in these conditions, listeners appear to have detected the target by detecting changes in the modulation depth between the masker-only and target+masker intervals, and spatial cues did not help in detecting these modulation changes. However, when the intensity rove was added in these conditions, the modulation-based colocated detection performance was impaired, while the spatially separated performance was not. Thus, spatial cues helped, bringing the spatially separated threshold to the no-rove levels, possibly by making it easier to use the modulation cue optimally.

When only the target was modulated, spatial cues provided a significant improvement in performance both when intensity was fixed across intervals in a trial (Experiment 1) and when intensity was roved (Experiment 2). For these stimuli, listeners were never given a good glimpse of the target, because the masker envelope was constant. In addition, the spatially separated thresholds were almost identical to the thresholds in the no-modulation and in-phase modulation conditions, and the size of the spatial benefit in the no-rove experiment was nearly identical to that in the no-modulation cue conditions. There are two possible explanations for the listeners' behavior in the target-only modulation condition when overall level was not roved. One possibility is that when the target and masker were colocated, listeners used an overall level to detect the target, and when target and masker were spatially separated, listeners used a spatial cue to detect the target. If so, then the modulation and spatial cues were subadditive in the target-only modulation case: listeners either used space or modulation. Alternatively, listeners may have used the modulation cue in the colocated target-only modulated condition and a combination of modulation and space cues in the spatially separated condition. If so, then spatial and modulation cues combined additively for this condition, but were combined subadditively in the masker-only and out-of-phase modulation conditions.

E. Final comments

After accounting for the better-ear acoustic benefit of spatial separation, the current study did not find any evidence for superadditive combination of modulation and space cues for detecting a broadband target embedded in a broadband masker. The results are consistent with two interpretations of the behavior when both cues were available and the level was fixed: (1) the subjects always used one of the cues, getting no benefit from the other one, or (2) the combination of modulation and space cues was additive when only the target was modulated, but the space cue contributed nothing to detection in the conditions in which the masker envelope was modulated and provided glimpses of the target. However, when the overall level was roved, spatial cues always helped performance when modulation was the main detection cue.

These results confirm the first of the proposed hypotheses (H1). The combined effect of modulation and spatial separation on detection is asymmetrical in that spatial separation improves detection performance more when the target is modulated and the masker is unmodulated than when the masker is modulated.

The results contradict our second hypothesis (H2). The combined effect of modulation and separation does not depend on the specific location of the target and masker, even though the contribution of binaural and across-frequency processing likely would vary in the different configurations. This result argues that the combined effect of modulation and spatial cues occurs at a stage that is later in the processing stream than the binaural processing occurring in the brainstem.

In contrast to the current stimuli, everyday auditory scenes contain objects that differ along many more dimensions than just their temporal envelopes and locations. It is difficult to extrapolate these findings to predict how modulation and spatial cues may interact for more complex stimuli. Nonetheless, it is likely that the main result, that modulation and space cues tend to contribute to detection subadditively, will also hold true for other stimuli differing in their spatial positions and modulation structure. However, it is also important to consider how our detection results compare to suprathreshold tasks, such as understanding speech embedded in fluctuating maskers. We find it intriguing that there is essentially no evidence for across-frequency integration in our experiments. In contrast, across-frequency integration is the basis of models that predict speech intelligibility in noise (e.g., see Zurek, 1993). We believe that the key difference between these results is that in our simpler detection task, any glimpse of the target (at any frequency) is sufficient for detection. In contrast, understanding speech requires the integration of information from different frequency bands and estimation of the absolute spectrotemporal content of the speech target. Thus, while the current results may be helpful in predicting how listeners detect a complex signal embedded in a competing fluctuating masker, they are only a first step in understanding how we analyze and understand the content of a complex signal in a setting containing multiple sound sources.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Institute on Deafness and Other Communication Disorders (5R01DC005778-03) and a grant from the Slovak Science Grant Agency (VEGA 1/3134/06). Rich Freyman gave a number of extraordinarily helpful suggestions as editor. In addition, two anonymous reviewers provided very helpful feedback on earlier versions of this work. The authors wish to thank H. Steven Colburn, Constantine Trahiotis, Les Bernstein, Bertrand Delgutte, Chris Mason, Eric Thompson, and Gin Best, for their helpful comments, and Jackie Jacobsen for help with data collection.

APPENDIX: LEARNING

Previous studies show that modulation detection performance improves with training over the course of hours (Wakefield and Viemeister, 1990; Dau and Ewert, 2004; Fitzgerald and Wright, 2005). In the current study, subjects did not receive extensive training prior to the experiment; each performed only one practice session in which thresholds for all conditions were measured once each (25 combinations of modulation and spatial configuration in Experiment 1 and 30 combinations in Experiment 2). To evaluate how learning influenced the results, data were analyzed as a function of the experimental session.

A three-way repeated-measure ANOVA was performed for both experiments on the data collapsed across masker locations [as in Fig. 3(a)], with factors of repeat (five levels), modulation type (five levels), and spatial separation (two levels). For Experiment 1, all two-way interactions were significant (repeat × modulation: $F_{16,96}=2.11$, $p=0.0134$; repeat × separation: $F_{4,24}=6.03$, $p=0.0017$; modulation × separation: $F_{4,24}=230$, $p<0.0001$), as were the main effects of modulation and separation ($p<0.0001$). For Experiment 2, the results were very similar (repeat × modulation: $F_{16,96}=1.69$, $p=0.062$; repeat × separation: $F_{4,24}=20.96$, $p<0.0001$; modulation × separation: $F_{4,24}=212$, $p<0.0001$; main effects of modulation and separation: $p<0.0001$). These results show that performance changes over time, and that the change depends on the specific combinations of modulation and of spatial separation.

Post hoc inspection reveals that the largest changes in SRM over time arose when only the target was modulated and when the target and masker were modulated out of phase. Panel A of Fig. 5 shows the thresholds for these conditions (target-only shown as squares; out-of-phase target and masker modulation shown as hexagrams), collapsed across the masker location and plotted as a function of the repeat, for both spatially colocated (open) and separated (filled) conditions. Panel B shows the SRM. The left-hand and right-hand panels show data from Experiments 1 and 2, respectively. Each symbol represents the across-subject mean (and within-subject 95% CI) of the thresholds obtained for one combination of repeat, spatial configuration, and modulation types.

Overall, TMR thresholds generally improved over time, as illustrated by the downward trend in all the graphs in panel A. However, a more detailed inspection shows that the size of this learning effect differed in the different conditions, and that these differences were consistent across the two experiments. When the stimuli were spatially separated, the target-only modulated thresholds (filled squares) improved by 2–3 dB over the five repeats, while the out-of-phase modulated thresholds (filled hexagrams) improved by 1 dB or less. On the other hand, when the stimuli were colocated, there was a roughly 3 dB improvement in the out-of-phase modulated thresholds (open hexagrams), while the improvement was negligible in the target-only modulated thresholds (open squares). As a result, the SRM tended to increase across sessions for target-only modulation stimuli

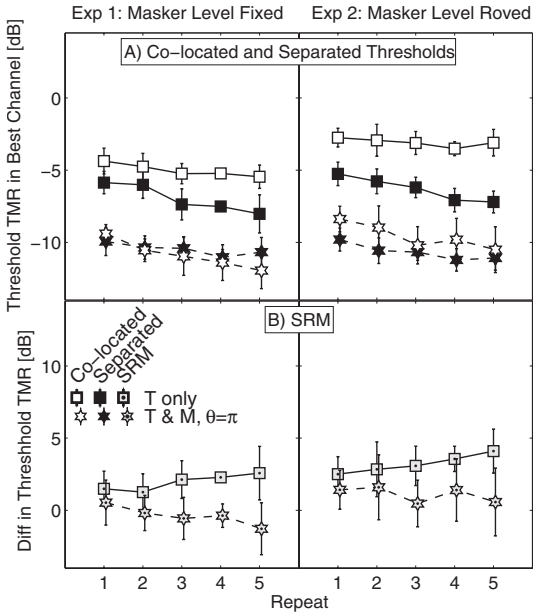

FIG. 5. Threshold TMR in the frequency-corrected best 1/3-octave channel (panel A) and the SRM (panel B) as a function of the measurement repeat in Experiment 1 (left-hand panels) and Experiment 2 (right-hand panels). Panel A: For each repeat, the data represent the across-subject mean (and within-subject 95% confidence interval) of the thresholds collapsed across the corresponding spatially separated or colocated conditions. Panel B: SRM, determined as the difference of the respective thresholds from Panel A.

but to decrease when the target and masker were modulated out of phase (panel B). Thus, while the SRMs for these two conditions differed by only about 1 dB in the first repeat, they differed by more than 4 dB by the fifth repeat.

At first glance, these changes seem difficult to understand. However, as discussed in the main text, spatial cues are generally not helpful for the out-of-phase conditions (hexagrams); in those conditions, performance is based on detecting (nonspatial) changes in modulation. The only effect of spatial cues in the out-of-phase modulation conditions was to make it easier to focus on this change in modulation (e.g., ignoring the distracting effects of intensity rove). Consistent with this, the main effect of learning in the out-of-phase modulation conditions is to improve how well listeners do when there are no spatial cues present and it is most difficult to focus attention on the modulation cue that underlies detection (open hexagrams).

In contrast, in the target-only modulation condition (squares), spatial cues provide a real advantage in target detection and allow detection at lower thresholds than when only monaural modulation and/or level cues are available. In these conditions, listeners improve most in their ability to use this subtle spatial cue (filled squares). However, listeners show little improvement in their ability to detect nonspatial changes in modulation or level with practice (open squares), perhaps because detection of modulation or detection of changes in level increases is a relatively simple task in which near-asymptotic performance is reached much faster (compared to the discrimination of modulation depth or detection of subtle spatial changes). As a result, SRM grows with time for the target-only modulation condition.

[1]The colocated spatial configuration with the target and masker at 90° was not measured in Experiment 1.

[2]In the case of a sinusoidal target, this correction can be computed by considering only the TMR change at the target frequency (Shinn-Cunningham *et al.*, 2005). If the relative contribution of each frequency to task performance is known for a broadband signal, the frequency-dependent TMR function can be used to predict performance (e.g., Zurek, 1993).

[3]This simple linear correction is purely phenomenological, rather than based on theoretical considerations. To the extent that this is the right correction to apply, it may reflect systematic deviations in the degree to which 1/3-octave filters approximate peripheral filtering as a function of frequency, differences in the internal noise of different frequency channels, or other systematic effects of frequency.

[4]Binaural and across-frequency processing may explain some of the dependence of the uncorrected thresholds on the masker locations. Specifically, in the spatially separated configuration of Fig. 2(a), the largest TMRs occur at low frequencies (below 2 kHz, full thick line) and the TMR profile in the right ear is relatively flat as a function of frequency. On the other hand, in the configurations of Figs. 2(b) and 2(c), the best frequency channel is at high frequencies and the TMRs vary significantly with frequency. These differences in the dominant spectral region suggest that binaural and across-frequency processing may contribute more to performance for the conditions of Fig. 2(a) than in the other two configurations, consistent with results in Figs. 1(b) and 1(e) (filled symbols in the *M* 0°, *T* 90° configuration are below the filled symbols for the other two configurations). However, while the binaural and across-frequency processing may explain why threshold TMRs tend to be lower when the masker is at 0° compared to the other configurations [leftmost versus middle and rightmost plots of Fig. 1(b)], they are not analyzed because (1) these factors cannot explain why some spatially separated thresholds are worse than the corresponding colocated thresholds in Figs. 1(b) and 1(e), and 2 the correction based on the frequency-dependent best-channel TMRs accounts for these differences, without considering binaural and across-frequency processing.

[5]Within-subject standard deviations are computed by subtracting out the mean performance (averaged across conditions) for each subject prior to the computation of variability. This method for computing variability is analogous to using subject as a factor in ANOVA analysis. In particular, the remaining variability shows how variable the across-condition results are after removing differences in overall performance across subjects. See the Appendix of Kopco *et al.* (2007) for further descriptions of this analysis.

[6]Comparisons of the current and previous results should be made with caution, as there are important differences in experimental procedures: for instance, none of the previous studies (Viemeister, 1979; Wakefield and Viemeister, 1990; Dau and Ewert, 2004; Dau, 1996) used the 40 Hz modulation frequency adopted in the present study. In addition, the current stimuli differ from the stimuli in the previous studies in their spectral content as they are filtered by the HRIRs.

ANSI (**1986**). *Specifications for Octave-Band and Fractional-Octave-Band Analog and Digital Filters, S1.1 (ASA 65-1986)* (American National Standards Institute, Inc., New York).

Asemi, N., Sugita, Y., and Suzuki, Y. (**2003**). "Auditory search asymmetry between pure tone and temporal fluctuating sounds distributed on the frontal-horizontal plane," Acta. Acust. Acust. **89**, 346–354.

Bernstein, L. R., and Trahiotis, C. (**1994**). "Detection of interaural delay in high-frequency sinusoidally amplitude-modulated tones, two-tone complexes, and bands of noise," J. Acoust. Soc. Am. **95**, 3561–3567.

Bernstein, L. R., and Trahiotis, C. (**2002**). "Enhancing sensitivity to interaural delays at high frequencies by using "transposed stimuli"," J. Acoust. Soc. Am. **112**, 1026–1036.

Buss, E., Hall, J. W., III, and Grose, J. H. (**2003**). "The masking level difference for signals placed in masker envelope minima and maxima," J. Acoust. Soc. Am. **114**, 1557–1564.

Buss, E., Hall, J. W., III, and Grose, J. H. (**2007**). "Individual differences in the masking level difference with a narrowband masker at 500 or 2000 Hz," J. Acoust. Soc. Am. **121**, 411–419.

Buus, S., Zhang, L., and Florentine, M. (**1996**). "Stimulus-driven, time-varying weights for comodulation masking release," J. Acoust. Soc. Am. **99**, 2288–2297.

Colburn, H. S. (**1977a**). "Theory of binaural interaction based on auditory-nerve data. II: Detection of tones in noise," J. Acoust. Soc. Am. **61**, 525–533; Colburn, H. S. (**1977b**). "Theory of binaural interaction based on auditory-nerve data. II: Detection of tones in noise. Supplementary material," J. Acoust. Soc. Am. AIP document no. PAPS JASMA-61–525–98.

Dau, T. (**1996**). "Modeling auditory processing of amplitude modulation," (Universität Oldenburg, Germany).

Dau, T., and Ewert, S. D. (**2004**). "External and internal limitations in amplitude-modulation processing," J. Acoust. Soc. Am. **116**, 478–490.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997**). "Modeling auditory processing of amplitude modulation I. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am. **102**, 2892–2905.

Durlach, N. I., Braida, L. D., and Ito, Y. (**1986**). "Towards a model for the discrimination of broadband stimuli," J. Acoust. Soc. Am. **80**, 63–72.

Fitzgerald, M. B., and Wright, B. A. (**2005**). "A perceptual learning investigation of the pitch elicited by amplitude-modulated noise," J. Acoust. Soc. Am. **118**, 3794–3803.

Freyman, R. L., Helfer, K. S., McCall, D. D., and Clifton, R. K. (**1999**). "The role of perceived spatial separation in the unmasking of speech," J. Acoust. Soc. Am. **106**, 3578–3588.

Gilkey, R. H., and Good, M. D. (**1995**). "Effects of frequency on free-field masking," Hum. Factors **37**, 835–843.

Good, M. D., Gilkey, R. H., and Ball, J. M. (**1997**). "The relation between detection in noise and localization in noise in the free field," in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. Gilkey and T. Anderson (Erlbaum, New York), pp. 349–376.

Green, D. M. (**1988**). *Profile Analysis. Auditory Intensity Discrimination* (Oxford University Press, New York).

Hall, J. W., III, Buss, E., and Grose, J. H. (**2006**). "Binaural comodulation masking release: Effects of masker interaural correlation," J. Acoust. Soc. Am. **120**, 3878–3888.

Hall, J. W., III, Haggard, M. P., and Fernandes, M. A. (**1984**). "Detection in noise by spectrotemporal pattern analysis," J. Acoust. Soc. Am. **76**, 50–56.

Herron, T. (**2005**). "C Language Exploratory Analysis of Variance with Enhancements," (Janunary 30, 2005 version. URL: http://www.ebire.org/hcnlab/software/cleave.html, date last viewed: May 19, 2008).

Kidd, G., Jr., Mason, C. R., Brantley, M. A., and Owen, G. A. (**1989**). "Roving-level tone-in-noise detection," J. Acoust. Soc. Am. **86**, 1310–1317.

Kopco, N. (**2005**). "Across-frequency integration in spatial release from masking," *Forum Acusticum* (OPAKFI, Budapest, Hungary), pp. 1607–1612.

Kopco, N., Best, V., and Shinn-Cunningham, B. G. (**2007**). "Sound localization with a preceding distractor," J. Acoust. Soc. Am. **121**, 420–432.

Kopco, N., and Shinn-Cunningham, B. G. (**2003**). "Spatial unmasking of nearby pure-tone targets in a simulated anechoic environment," J. Acoust. Soc. Am. **114**, 2856–2870.

Lane, C., Kopco, N., Delgutte, B., Shinn-Cunningham, B., and Colburn, H. (**2004**). "A cat's cocktail party: Psychophysical, neurophysiological, and computational studies of spatial release from masking," in *Auditory Signal Processing: Physiology, Psychoacoustics, and Models*, edited by D. Pressnitzer, A. d. Cheveigne, S. McAdams, and L. Collet (Springer, Dourdan, France), pp. 405–413.

Lane, C. C., and Delgutte, B. (**2005**). "Neural correlates and mechanisms of spatial release from masking: Single-unit and population responses in the inferior colliculus," J. Neurophysiol. **94**, 1180–1198.

Levitt, H. (**1971**). "Transformed up-down methods in psychophysics," J. Acoust. Soc. Am. **49**, 467–477.

Mason, C. R., Kidd, G., Hanna, T. E., and Green, D. M. (**1984**). "Profile analysis and level variation," Hear. Res. **13**, 269–275.

Moore, B. C. J. (**2003**). *An Introduction to the Psychology of Hearing*, 5th ed. (Academic, San Diego, CA).

Saberi, K., Dostal, L., Sadralodabai, T., Bull, V., and Perrott, D. R. (**1991**). "Free-field release from masking," J. Acoust. Soc. Am. **90**, 1355–1370.

Sheft, S., and Yost, W. A. (**1997**). "Binaural modulation detection interference," J. Acoust. Soc. Am. **102**, 1791–1798.

Shinn-Cunningham, B. G., Ihlefeld, A., Satyavarta, and Larson, E. (**2005**). "Bottom-up and top-down influences on spatial unmasking," Acta. Acust. Acust. **91**, 967–979.

Shinn-Cunningham, B. G., Kopco, N., and Martin, T. J. (**2005**). "Localizing nearby sound sources in a classroom: Binaural room impulse responses," J. Acoust. Soc. Am. **117**, 3100–3115.

Spiegel, M. F., Picardi, M. C., and Green, D. M. (**1981**). "Signal and masker

uncertainty in intensity discrimination," J. Acoust. Soc. Am. **70**, 1015–1019.
Stellmack, M. A., Viemeister, N. F., and Byrne, A. J. (**2005**). "Monaural and interaural temporal modulation transfer functions measured with 5-kHz carriers," J. Acoust. Soc. Am. **118**, 2507–2518.
Stellmack, M. A., Viemeister, N. F., and Byrne, A. J. (**2006**). "Discrimantion of depth of sinusoidal amplitude modulation with and without roved carrier levels (L)," J. Acoust. Soc. Am. **119**, 37–40.
Sterbing, S. J., D'Angelo, W. R., Ostapoff, E.-M., and Kuwada, S. (**2003**). "Effects of amplitude modulation on the coding of interaural time differences of low-frequency sounds in the inferior colliculus. I. Response properties," J. Neurophysiol. **90**, 2818–2826.
Ury, H. K., and Wiggins, A. D. (**1971**). "Large sample and other multiple comparisons among means," Br. J. Math. Stat. Psychol. **24**, 174–194.
van de Par, S., and Kohlrausch, A. (**1997**). "A new approach to comparing binaural masking level differences at low and high frequencies," J. Acoust. Soc. Am. **101**, 1671–1680.
van de Par, S., and Kohlrausch, A. (**1998**). "Comparison of monaural (CMR) and binaural (BMLD) masking release," J. Acoust. Soc. Am. **103**, 1573–1579.
van de Par, S., Kohlrausch, A., Breebaart, J., and McKinney, M. (**2004**). "Discrimination of different temporal envelope structures of diotic and dichotic target signals within diotic wide-band noise," in *Auditory Signal Processing: Physiology, Psychoacoustics, and Models*, edited by D. Pressnitzer, A. de Cheveigné, S. McAdams, and L. Collet (Springer, New York), pp. 398–404.
Viemeister, N. F. (**1979**). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am. **66**, 1364–1380.
Wakefield, G. H., and Viemeister, N. F. (**1990**). "Discrimination of modulation depth of sinusoidal amplitude modulation (SAM) noise," J. Acoust. Soc. Am. **88**, 1367–1373.
Widin, G. P., Viemeister, N. F., and Bacon, S. P. (**1986**). "Effects of forward and simultaneous masking on intensity discrimination," J. Acoust. Soc. Am. **80**, 108–111.
Winter, I. M., Neuert, V., and Verhey, J. L. (**2004**). "Comodulation masking release and the role of wideband inhibition in the cochlear nucleus," in *Auditory Signal Processing: Physiology, Psychoacoustics, and Models*, edited by D. Pressnitzer, A. de Cheveigné, S. McAdams, and L. Collet (Springer, New York), pp. 321–327.
Wojtczak, M., and Viemeister, N. F. (**2005**). "Forward masking of amplitude modulation: Basic characteristics," J. Acoust. Soc. Am. **118**, 3198–3210.
Zurek, P. M. (**1993**). "Binaural advantages and directional effects in speech intelligibility," in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. Studebaker and I. Hochberg (College-Hill Press, Boston, MA).
Zurek, P. M., Freyman, R. L., and Balakrishnan, U. (**2004**). "Auditory target detection in reverberation," J. Acoust. Soc. Am. **115**, 1609–1620.

Spatial unmasking of nearby speech sources in a simulated anechoic environment

Barbara G. Shinn-Cunningham[a)]
Boston University Hearing Research Center, Departments of Cognitive and Neural Systems and Biomedical Engineering, Boston University, 677 Beacon St., Room 311, Boston, Massachusetts 02215

Jason Schickler
Boston University Hearing Research Center, Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215

Norbert Kopčo
Boston University Hearing Research Center, Department of Cognitive and Neural Systems, Boston University, Boston, Massachusetts 02215

Ruth Litovsky
Boston University Hearing Research Center, Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215

(Received 18 August 2000; revised 24 May 2001; accepted 25 May 2001)

Spatial unmasking of speech has traditionally been studied with target and masker at the same, relatively large distance. The present study investigated spatial unmasking for configurations in which the simulated sources varied in azimuth and could be either near or far from the head. Target sentences and speech-shaped noise maskers were simulated over headphones using head-related transfer functions derived from a spherical-head model. Speech reception thresholds were measured adaptively, varying target level while keeping the masker level constant at the ''better'' ear. Results demonstrate that small positional changes can result in very large changes in speech intelligibility when sources are near the listener as a result of large changes in the overall level of the stimuli reaching the ears. In addition, the difference in the target-to-masker ratios at the two ears can be substantially larger for nearby sources than for relatively distant sources. Predictions from an existing model of binaural speech intelligibility are in good agreement with results from all conditions comparable to those that have been tested previously. However, small but important deviations between the measured and predicted results are observed for other spatial configurations, suggesting that current theories do not accurately account for speech intelligibility for some of the novel spatial configurations tested.
[DOI: 10.1121/1.1386633]

PACS numbers: 43.66.Pn, 43.66.Ba, 43.71.An, 43.66.Rq [LRB]

I. INTRODUCTION

When a target of interest (T) is heard concurrently with an interfering sound (a ''masker,'' M), the locations of both target and masker have a large effect on the ability to detect and perceive the target. Previous studies have examined how T and M locations affect performance in both detection (e.g., see the review in Durlach and Colburn, 1978 or, for example, recent work such as Good, Gilkey, and Ball, 1997) and speech intelligibility tasks (e.g., see the recent review by Bronkhorst, 2000). Generally speaking, when the T and M are located at the same position, the ability to detect or understand T is greatly affected by the presence of M; when either T or M is displaced, performance improves.

While there are many studies of spatial unmasking for speech (e.g., see Hirsh, 1950; Dirks and Wilson, 1969; MacKeith and Coles, 1971; Plomp and Mimpen, 1981; Bronkhorst and Plomp, 1988; Bronkhorst and Plomp, 1990; Peissig and Kollmeier, 1997; Hawley, Litovsky, and Colburn, 1999), all of the previous studies examined targets and maskers that were located far from the listener. These studies examined spatial unmasking as a function of angular separation of T and M without considering the effect of distance. One goal of the current study was to measure spatial unmasking for a speech reception task when a speech target and a speech-shaped noise masker are within 1 meter of the listener. In this situation, changes in source location can give rise to substantial changes in both the overall level and the binaural cues in the stimuli reaching the ears (e.g., see Duda and Martens, 1997; Brungart and Rabinowitz, 1999; Shinn-Cunningham, Santarelli, and Kopčo, 2000). Because the acoustics for nearby sources can differ dramatically from those of more distant sources, insights gleaned from previous studies may not apply in these situations. In addition, previous models (which do a reasonably good job of predicting performance on similar tasks; e.g., see Zurek, 1993) may not be able to predict what occurs when sources are close to the listener precisely because the acoustic cues at the ears are so different than those that arise for relatively distant sources.

For noise maskers that are statistically stationary (such

a)Electronic mail: shinn@cns.bu.edu

 0001-4966/2001/110(2)/1118/12/$18.00

as steady-state broadband noise in anechoic settings, but not, for instance, amplitude-modulated noise or speech maskers), spatial unmasking can be predicted from simple changes in the acoustic signals reaching the ears (e.g., see Bronkhorst and Plomp, 1988; Zurek, 1993). For T fixed directly in front of a listener, lateral displacement of M causes changes in (1) the relative level of the T and M at the ears (i.e., the target to masker level ratio, or TMR), which will differ at the two ears (a monaural effect) and (2) the interaural differences in T compared to M (a binaural effect, e.g., see Zurek, 1993). For relatively distant sources, the first effect arises because the level of the masker reaching the farther ear decreases (particularly at moderate and high frequencies) as the masker is displaced laterally (giving rise to the acoustic ''head shadow''). Thus, as M is displaced from T, one of the two ears will receive less energy from M, resulting in a ''better-ear advantage.'' Also, for relatively distant sources the most important binaural contribution to unmasking occurs when T and M give rise to different interaural time differences (ITDs), resulting in differences in interaural phase differences (IPDs) in T and M, at least at some frequencies (e.g., see Zurek, 1993). The overall size of the release from masking that can be obtained when T is located in front of the listener and a steady-state M is laterally displaced (and both are relatively distant from the listener) is on the order of 10 dB (e.g., see Plomp and Mimpen, 1981; Bronkhorst and Plomp, 1988; Peissig and Kollmeier, 1997; Bronkhorst, 2000). Of this 10 dB, roughly 2–3 dB can be attributed to binaural processing of IPDs, with the remainder resulting from head shadow effects (e.g., see Bronkhorst, 2000).

If one restricts the target and masker to be at least 1 meter from the listener, the only robust effect of distance on the stimuli at the ears is a change in overall level (e.g., see Brungart and Rabinowitz, 1999). Thus, for relatively distant sources, the effect of distance can be predicted simply from considering the dependence of overall target and masker level on distance; there are no changes in binaural cues, the better-ear-advantage, or the difference in the TMR at the better and worse ears.

There are important differences between how the acoustic stimuli reaching the ears change when a sound source is within a meter of and when a source is more than a meter from the listener (e.g, see Duda and Martens, 1997; Brungart and Rabinowitz, 1999; Shinn-Cunningham *et al.*, 2000). For instance, a small displacement of the source towards the listener can cause relatively large increases in the levels of the stimuli at the ears. In addition, for nearby sources, the interaural level difference (ILD) varies not only with frequency and laterality but also with source distance. Even at relatively low frequencies, for which naturally occurring ILDs are often assumed to be zero (i.e., for sources more than about a meter from the head), ILDs can be extremely large. In fact, these ILDs can be broken down into the traditional ''head shadow'' component, which varies with direction and frequency, and an additional component that is frequency independent and varies with source laterality and distance (Shinn-Cunningham *et al.*, 2000).

In the ''distant'' source configurations previously studied, the better ear is only affected by the relative laterality of T versus M; the only spatial unmasking that can arise for T and M in the same direction is a result of equal overall level changes in the stimuli at the two ears. Moving T closer than M will improve the SRT while moving T farther away will decrease performance, simply because the level of the target at both ears varies with distance (equivalently). In contrast, when a source is within a meter of the head, the relative level of the source at the two ears depends on distance. Changing the distance of T or M can lead not only to changes in overall energy, but changes in the amount of unmasking that can be attributed to binaural factors, the difference in the TMR at the two ears (as a function of frequency), and even which is the better ear. In addition, overall changes in the level at the ears can be very large, even for small absolute changes in distance. Although the distances for which these effects arise are small, in a real ''cocktail party'' it is not unusual for a listener to be within 1 meter of a target of interest (i.e., in the range for which these effects are evident).

We are aware of only one previous study of spatial unmasking for speech intelligibility in which large ILDs were present in both T and M (Bronkhorst and Plomp, 1988). In this study, the total signal to one ear was attenuated in order to simulate monaural hearing impairment. Unlike the Bronkhorst and Plomp study, the current study focuses on the spatial unmasking effects that occur when realistic combinations of IPD and ILD, consistent with sources within 1 m of the listener, are simulated for different T and M geometries.

II. EXPERIMENTAL APPROACH

A common measure used to assess spatial unmasking effects on speech tasks is the speech reception threshold (SRT), or the level at which the target must be presented in order for speech intelligibility to reach some predetermined threshold level. The amount of spatial unmasking can be summarized as the difference (in dB) between the SRT for the target/masker configuration of interest and the SRT when T and M are located at the same position.

In these experiments, SRT was measured for both ''nearby'' sources (15 cm from the center of the listener's head) and ''distant'' sources (1 m from the listener). Tested conditions included those in which (1) the speech target was in front of the listener and M was displaced in angle and distance; (2) M was in front of the listener and T displaced in angle and distance; and (3) T and M were both located on the side, but T and M distances were manipulated.

The goals of this study were to (1) measure how changes in spatial configuration of T and M affect SRT for sources near the listener; (2) explore how the interaural level differences that arise for nearby sources affect spatial unmasking; and (3) quantify the changes in the acoustic cues reaching the two ears when T and/or M are near the listener.

A. Subjects

Four healthy undergraduate students (ages ranging from 19–23 years) performed the tests. All subjects had normal hearing thresholds (within 15 dB HL) between 250 and 8000 Hz as verified by an audiometric screening. All subjects were native English speakers. One of the subjects was author JS

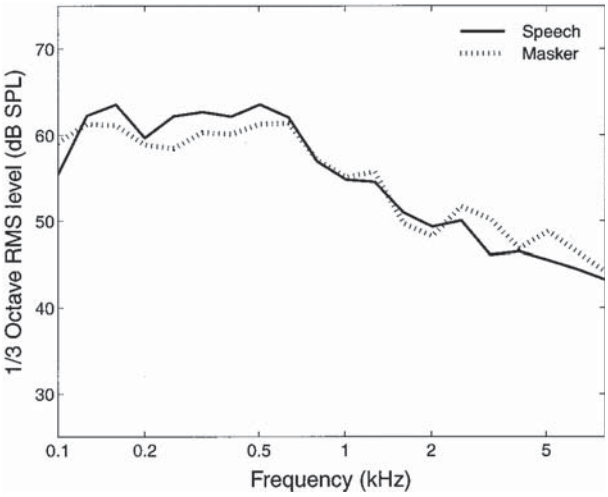

FIG. 1. Average spectral shape of speech-shaped noise masker and speech targets, prior to HRTF processing.

with relatively little experience in psychoacoustic experiments; the other three subjects were naive listeners with no prior experience.

B. Stimuli

1. Source characteristics

In the experiments, the target (T) consisted of a high-context sentence selected from the IEEE corpus (IEEE, 1969). Sentences were chosen from 720 recordings made by two different male speakers. These materials have been employed previously in similar speech intelligibility experiments (Hawley *et al.*, 1999). The recordings, ranging from 2.41–3.52 s in duration, were scaled to have the same rms pressure value in their ''raw'' (nonspatialized) forms. An example sentence is ''The DESK and BOTH CHAIRS were PAINTED TAN,'' with capitalized words representing ''key words'' that are scored in the experiment (see Sec. C).

The masker (M) was speech-shaped noise generated to have the same spectral shape as the average of the speech tokens used in the study. For each masker presentation, a random 3.57-s sample was taken from a long (24-s) sample of speech-shaped noise (this length guaranteed that all words in all sentences were masked by the noise). Figure 1 shows the rms pressure level in 1/3-octave bands (dB SPL) of the 24-s-long masking noise and the average of the spectra of the speech samples used in the study.

2. Stimulus generation

Raw digital stimuli (i.e., IEEE sentences and speech-shaped noise sampled at 20 kHz) were convolved with spherical-head head-related transfer functions (HRTFs) off-line (see below). T and M were then scaled (in software) to the appropriate level for the current configuration and trial. The resulting binaural T and M were then summed in software and sent to Tucker-Davis Technologies (TDT) hardware to be converted into acoustic stimuli (using the same equipment setup described in Hawley *et al.*, 1999). Digital signals were processed through left- and right-channel D/A converters (TDT DD3-8), low-pass filters (10-kHz cutoff; TDT FT5), and attenuators (TDT PA4). The resulting binaural analog signals were passed through a Tascam power amplifier (PA-20 MKII) connected to Sennheiser headphones (HD 520 II). No compensation for the headphone transfer function was performed. A personal computer (Gateway 2000 486DX) controlled all equipment and recorded results.

3. Spatial cues

In order to simulate sources at different positions around the listener, spherical-head HRTFs were generated for all the positions from which sources were to be simulated. These HRTFs were generated using a mathematical model of a spherical (9-cm-radius) head with diametrically opposed point receivers (ears; for more details about the model or traits of the resulting HRTFs see Rabinowitz *et al.*, 1993; Brungart and Rabinowitz, 1999; Shinn-Cunningham *et al.*, 2000). Source stimuli (T and M) were convolved to generate binaural signals similar to those that a listener would experience if the T and M were played from specific positions in anechoic space.

It should be noted that the spherical-head HRTFs are not particularly realistic. They contain no pinnae cues (i.e., contain no elevation information), are more symmetrical than true HRTFs, and are not tailored to the individual listener. As a result, sources simulated from these HRTFs are distinguishably different from sounds that would be heard in a real-world anechoic space. As a result, the sources simulated with these HRTFs may not have been particularly ''externalized,'' although they were generally localized at the simulated direction. There was no attempt to evaluate the realism, externalization, or localizability of the simulated sources using the spherical-head HRTFs. Nonetheless, the spherical-head HRTFs contain all the acoustic cues that are unique to sources within 1 m of the listener (i.e., large ILDs that depend on distance, direction, and frequency; changes in IPD with changes in distance), a result confirmed by comparisons with measurements of human subject and KEMAR HRTFs for sources within 1 m (see, for example, Brown, 2000; Shinn-Cunningham, 2000). Further, because the unique acoustic attributes that arise for free-field near sources are captured in these HRTFs, we believe that any unique behavioral consequences of listening to targets and maskers that are near the listener will be observed in these experiments.

4. Spatial configurations

In different conditions, the target and masker were simulated from any of six locations in the horizontal plane containing the ears; that is, at three azimuths (0°, 45°, and 90° to the right of midline) and two distances from the center of the head (15 cm and 1 m). The 15 spatial configurations investigated in this study are illustrated in Fig. 2. The three panels depict three different conditions: target location fixed at (0°, 1 m) [Fig. 2(a)], masker fixed at (0°, 1 m) [Fig. 2(b)] and target and masker both at 90° [Fig. 2(c)]. All subsequent graphs are arranged similarly. Note that the configuration in which T and M are both located at (0°, 1 m) appears in both panels (a) and (b) of Fig. 2; this spatial configuration was the (diotic) reference used in computing spatial masking effects.

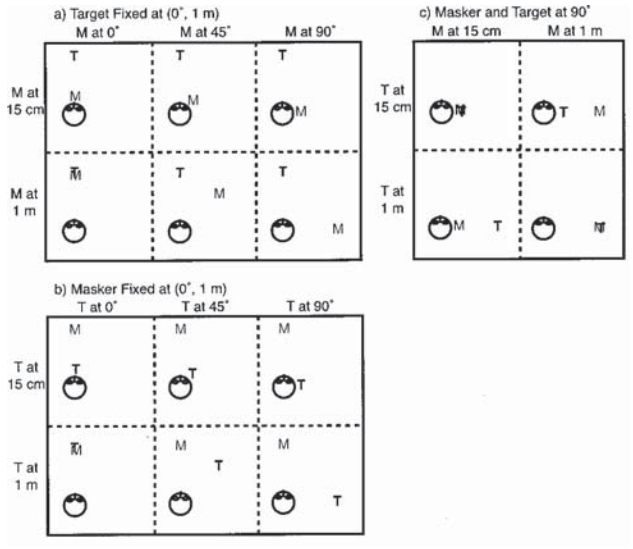

FIG. 2. Spatial configurations of target (T) and masker (M). Conditions: (a) T fixed (0°, 1 m); (b) M fixed (0°, 1 m); and (c) T and M at 90°.

5. Presentation level

If we had simulated a masking source emitting the same energy from different distances and directions, the level of the masker reaching the better ear would vary dramatically with the simulated position of M. In addition, depending on the location of M, the better ear can be either the ear nearer or farther from T. For instance, if T is located at (90°, 1 m) and M is located at (90°, 15 cm) [see Fig. 2(c), bottom left panel], T is nearer to the right ear, but the left ear will be the ''better ear.''

In order to roughly equate the masker energy *reaching the better ear* (as opposed to keeping constant the distal energy of the simulated masker), masker level was normalized so that the root-mean-square (rms) pressure of M at the better ear was always 72 dB SPL. With this choice, the masker was always clearly audible at the worse ear (even when the masker level was lower at the worse ear) and at a comfortable listening level at the worse ear (even when the masker level was higher at the worse ear). Of course, the worse-ear masker level varied with spatial configuration, and could either be greater or less than 72 dB SPL depending on the locations of T and M.

C. Experimental procedure

All experiments were performed in a double-walled sound-treated booth in the Binaural Hearing Laboratory of the Boston University Hearing Research Center.

An adaptive procedure was used to estimate the SRT for each spatial configuration of T and M. In each adaptive run, the T level was adaptively varied to estimate the SRT, which was defined as the level at which subjects correctly identified 50% of the T sentence key words.

For each configuration, at least three independent, adaptive-run threshold estimates were averaged to form the final threshold estimate. If the standard error in the repeated measures was greater than 1 dB, additional adaptive runs were performed until the standard error in this final average was equal to or less than 1 dB.

The T and M locations were not known *a priori* by the subject, but were held constant through a run, which consisted of ten trials. Runs were ordered randomly and broken into sessions consisting of approximately seven runs each.

Within a run, the first sentence of each block was repeated multiple times in order to set the T level for subsequent trials. The first sentence in each run was first played at 44 dB SPL in the better ear. The sentence was played repeatedly, with its intensity increased by 4 dB with each repetition, until the subject indicated (by subjective report) that he could hear the sentence. The level at which the listener reported understanding the initial sentence set the T level for the second trial in the run. On each subsequent trial, a new sentence was presented to the subject. The subject typed in the perceived sentence on a computer keyboard. The actual sentence was then displayed (along with the subject's typed response) on a computer monitor (visible to the subject) with five ''key words'' capitalized. The subject then counted up and entered into the computer the number of correct key words perceived. Scoring was strict, with incorrect suffixes scored as ''incorrect;'' however, homophones and misspellings were not penalized. Listeners heard only one presentation of each T sentence.

If the subject identified at least three of the five key words correctly, the level of the T was decreased by 2 dB on the subsequent trial. Otherwise (i.e., if the subject identified two or fewer key words), the level of the T was increased by 2 dB. Thus, if the subject performed at or above 60% correct, the task was made more difficult; if the subject performed at or below 40% correct, the task was made easier. This procedure (which, in the limit, will converge to the presentation level at which the subject will achieve 50% correct) was repeated until ten trials were scored. SRT was estimated as the average of the presentation levels of the T on the last eight (of ten) trials.

III. RESULTS

A. Target-to-masker levels at speech reception threshold

In order to visualize the changes in relative spectral levels of T and M with spatial configuration, the average TMR in third-octave spectral bands was computed as a function of center frequency at 50%-correct SRT and plotted in Fig. 3.

By construction (because T and M have the same spectral shape), the TMR is equal in both ears and independent of frequency for configurations in which T and M are located at the same position (i.e., for two diotic configurations and two configurations with T and M at 90°). However, in general, the overall spectral shape of both T and M depends on spatial configuration and the TMR varies with frequency.

In the diotic reference configuration, the TMR is −7.6 dB [e.g., see Fig. 3(a), bottom left panel]. In other words, when the diotic sentence is presented at a level 7.6 dB below the diotic speech-shaped noise, subjects achieve threshold performance in the reference configuration. This diotic reference TMR is plotted as a dashed horizontal line in all panels

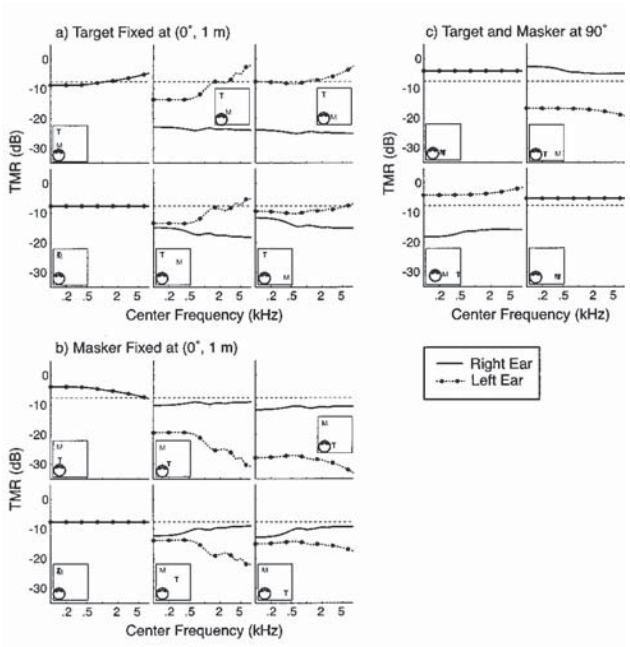

FIG. 3. Target-to-masker level ratio (TMR) in 1/3-octave frequency bands for left (dotted lines with symbols) and right (solid lines) ears as a function of center frequency at speech reception threshold. Conditions: (a) T fixed (0°, 1 m); (b) M fixed (0°, 1 m); and (c) T and M at 90°.

in order to make clear how the TMR varies with spatial configuration. When threshold TMR at the better ear is lower than the diotic reference TMR, the results indicate the presence of spatial masking effects that cannot be explained by overall level changes. In such cases, other factors, such as differences in binaural cues in T and M, are likely to be responsible for the improvements in SRT.

Figure 3(a) shows the results when T is fixed at (0°, 1 m). For these spatial configurations, the TMR at the better (left) ear (dotted line with symbols) is generally equal to or smaller than the reference TMR. TMR is lowest when M is located at (45°, 1 m) (bottom center panel); in this case, the TMR at low frequencies is as much as 14 dB below the diotic reference TMR (the TMR at higher frequencies is approximately equal to the diotic reference TMR). The worse-ear TMR (right ear; solid line) is often much smaller than that of the better ear, particularly when M is at 15 cm.

When the masker is fixed at the reference position (0°, 1 m) [Fig. 3(b)], the TMR at the better (right) ear (solid line) is below the reference TMR at all frequencies for all four cases in which T is laterally displaced. The magnitude of this improvement is roughly the same (2–3 dB) whether T is near or far, at 45° or 90°. In the diotic case for which T is at (0°, 1 m) and M is at (0°, 15 cm) [top-left panel in Fig. 3(b)], the TMR is roughly 4 dB larger than in the diotic reference configuration. This result indicates a small spatial *dis*-advantage in this diotic configuration compared to the ''typical'' diotic reference configuration when T and M are both distant after taking into account the overall level of M.

In all four configurations for which both T and M are located laterally [Fig. 3(c)], the TMR at the better ear is roughly 3–4 dB larger at all frequencies than the diotic reference TMR. In other words, listeners need a laterally located speech source to be presented at a relatively high level when it competes with a masker located in the same lateral direction. This is even true when M is at 1 m and T is at 15 cm [top right panel of Fig. 3(c)], despite the fact that the better- (right-) ear stimulus is at a substantially higher overall level than the worse- (left-) ear stimulus in this configuration.

B. Mean difference in monaural TMRs

The results in Fig. 3 show that the difference in the TMRs at the two ears can be very large when either T or M is near the listener (a direct consequence of the very large ILDs that arise for these sources). This difference is important for understanding and quantifying the advantage of having two ears, independent of any binaural processing advantage. For instance, if a monaurally impaired listener's intact ear is the acoustically worse ear, the impaired listener will be at a larger disadvantage for many of the tested configurations than when both T and M are distant. In order to quantify the magnitude of these acoustic effects, the absolute value of the mean of the difference in left- and right-ear TMR was calculated, averaged across frequencies up to 8000 Hz.

The leftmost data column in Table I gives the mean of $|\mathrm{TMR}_{\mathrm{right}} - \mathrm{TMR}_{\mathrm{left}}|$ at SRT, averaged across frequency. Because the TMRs change with frequency, this estimate cannot predict SRT directly; for instance, moderate frequencies (e.g., 2000–5000 Hz) convey substantially more speech information than lower frequencies. Nonetheless, these calculations give an objective, acoustic measure, weighting all frequencies equally, of differences in the better and worse ear signals.

From symmetry and because T and M have the same spectral shape, the difference in better- and worse-ear TMR is the same if M is held at (0°, 1 m) and T is moved or T is fixed and M is moved (see Table I, comparing top and center sections).

For configurations in which both T and M are far from the head, the acoustic difference in the TMRs at the two ears ranges from 5–10 dB, depending on the angular separation of T and M. If T remains fixed and a laterally located M is moved from 1 m to 15 cm (or vice versa), the difference between the better and worse ear TMR increases substantially. For instance, with T fixed at (0°, 1 m) and M at (90°, 15 cm), the difference in TMR is nearly 20 dB (third line in Table I). For spatial configurations in which one source is near the head but not in the median plane, part of this difference in better- and worse-ear TMR arises from ''normal'' head-shadow effects and part arises due to differences in the relative distance from the source to the two ears (Shinn-Cunningham *et al.*, 2000).

In the configurations for which both T and M are located at 90°, there is no difference in the TMR at the ears when T and M are at the same distance. When one source is near and one is far, the TMR at the ears differs by roughly 13 dB.

It should be noted that there are even more extreme spatial configurations than those tested here. For instance, with T at (−90°, 15 cm) and M at (+90°, 15 cm) the acoustic difference in the TMRs at the two ears would be on the order of 40 dB (i.e., twice the difference obtained when one

TABLE I. Spatial effects for different spatial configurations tested. Leftmost data column shows the mean of the absolute difference $|TMR_{right} - TMR_{left}|$ at SRT, averaged across frequencies up to 8000 Hz. The second data column gives the predicted magnitude of the difference in the monaural left- and right-ear SRTs from the Zurek model calculations. The third data column gives the binaural advantage calculated from Zurek model calculations (the difference in predicted SRT for binaural and monaural better-ear listening conditions).

			Left/right asymmetry (acoustic analysis) (dB)	Left/right asymmetry (Zurek predictions) (dB)	Binaural advantage (Zurek predictions) (dB)
T (0°, 1 m)	M (15 cm)	M (0°)	0	0	0
		M (45°)	17.5	14.6	2.0
		M (90°)	19.6	17.9	1.5
	M (1 m)	M (0°)	0	0	0
		M (45°)	9.8	7.5	2.4
		M (90°)	6.4	5.2	2.2
M (0°, 1 m)	T (15 cm)	T (0°)	0	0	0
		T (45°)	17.5	14.5	1.5
		T (90°)	19.6	17.2	1.5
	T (1 m)	T (0°)	0	0	0
		T (45°)	9.8	7.5	1.9
		T (90°)	6.4	5.2	2.2
T & M (90°)	T (15 cm)	M (15 cm)	0	0	0
		M (1 m)	13.2	12.6	0.8
	T (1 m)	M (15 cm)	13.2	12.6	0.9
		M (1 m)	0	0	0

source is diotic and one source is at 90°, 15 cm). This analysis demonstrates that one novel outcome of T and M being very close to the head is that the difference in the TMRs at the two ears can be dramatically larger than in previously tested configurations.

C. Spatial unmasking

Figure 4 plots the amount of spatial unmasking for each spatial configuration.[1] In the figure, the amount of "spatial unmasking" equals the decrease in the distal energy the target source must emit for subjects to correctly identify 50% of the target key words if the distal energy emitted by the masking source were held constant. This analysis includes changes in the overall level of T and M reaching the ears with changes in source position (and assumes that SRT depends only on TMR and is independent of the absolute level of the masker for the range of levels considered).

When T is fixed at (0°, 1 m) [Fig. 4(a)], the release from masking is largest when the 1-m M is at 45° and decreases slightly when M is at 90°. The dependence of the unmasking on M distance is roughly the same for all M directions: moving M from 1 m to 15 cm increases the required T level by roughly 13 dB for M in all tested directions (0°, 45°, and 90°).

When M is fixed ahead [Fig. 4(b)], moving the 1-m-distant T to either 45° or 90° results in the same unmasking. Moving the T close to the head (15 cm) results in a large amount of spatial unmasking, primarily due to increases in the level of T reaching the ears. For a given T direction, the effect of decreasing the distance of T increases with its lateral angle.

Figure 4(c) shows the spatial unmasking that arises when T and M are both located at 90°. When T and M are at the same distance [either at 15 cm, circles at left of Fig. 4(c); or at 1 m, squares at right of Fig. 4(c)], there is a 3-dB increase in the level the target source must emit compared to the reference configuration. When T and M are at different distances, spatial unmasking results are dominated by differences in the relative distances to the head.

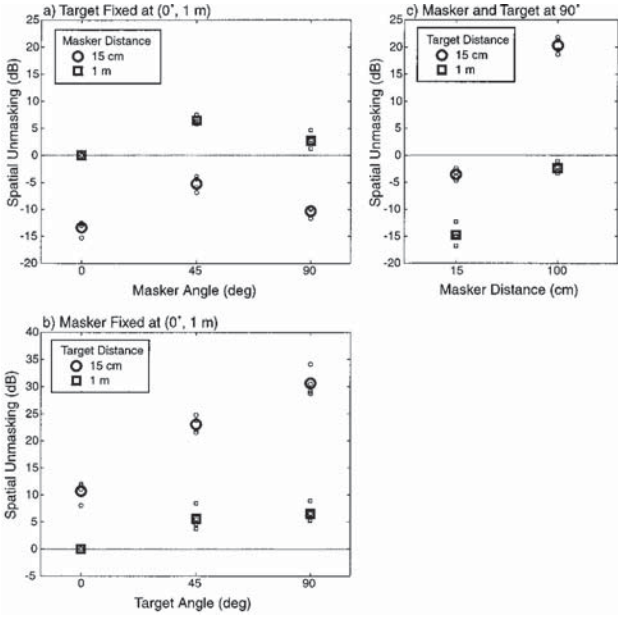

FIG. 4. Spatial advantage (energy a target emits at threshold for a constant-energy masker) relative to the diotic configuration. Positive values are decreases in emitted target energy. Large symbols give the across-subject mean; small symbols show individual subject results. Conditions: (a) T fixed (0°, 1 m); (b) M fixed (0°, 1 m); and (c) T and M at 90°.

D. Discussion

Our findings are generally consistent with previous results that show that speech intelligibility improves when T and M give rise to different IPDs, and that spatially separating a masker and target tends to reduce threshold TMR.

However, in some of the spatial configurations tested, the threshold TMR at the better ear is greater than the TMR in the diotic reference configuration. For instance, in all four spatial configurations with T and M at 90° [Fig. 3(c)], the better-ear TMR is roughly the same (independent of the relative levels of the better and worse ears) and elevated compared to the TMR in the diotic reference configuration. These results are inconsistent with predictions from previous models, which generally assume that binaural performance is always at least as good as would be observed if listeners were presented with the better-ear stimulus monaurally. Discrepancies between the current findings and predictions from an existing model (Zurek, 1993) are considered in detail in the next section.

For distant sources, changing the distance of T or M may change the overall level at the better ear, but it causes an essentially identical change at the worse ear. Thus, the difference between listening with the worse and the better ears is independent of T and M distance when T and M are at least 1 m from the listener. One of the novel effects that arises when either T or M is within 1 meter of the head is that the difference between the TMR at the better and worse ears can be dramatically larger than if both T and M are distant (see Table I). For the configurations tested, the difference in the TMRs at the two ears can be nearly double the difference that occurs when both T and M are at least a meter from the listener [e.g., 19.6 dB for a diotic T and M at (90°, 15 cm) versus 9.8 dB for diotic T and M at (90°, 1 m)].

Analysis of the spatial unmasking (Fig. 4) emphasizes the large changes in overall level that can arise with small displacements of a source near the listener. For the configurations tested, the change in the level that the target must emit to be intelligible against a constant level masker ranges from −31 to +15 dB (relative to the diotic reference configuration).

IV. MODEL PREDICTIONS

A. Zurek model of spatial unmasking of speech

Zurek (1993) developed a model based on the Articulation Index (AI,[2] Fletcher and Galt, 1950; ANSI, 1969; Pavlovic, 1987) to predict speech intelligibility as a function of target and masker location. AI is typically computed for a single-channel system as a weighted sum of target-to-masker ratios (TMRs) across third-octave frequency bands. In Zurek's model, the TMRs at both ears are considered, along with interaural differences in the T and M.

To compute the predicted intelligibility, Zurek's model first computes the actual TMR at each ear in each of 15 third-octave frequency bands (spaced logarithmically between 200 to 5000 Hz). The "effective TMR" (R_i) in each frequency band i is the sum of (1) the larger of the two true TMRs at the left and right ears and (2) an estimate of the "binaural advantage" in band i. The binaural advantage in each band, derived from a simplified version of Colburn's model of binaural interaction (Colburn, 1977a, b), depends jointly on center frequency and the relative IPD of target and masker at the center frequency of the band. The advantage in a particular frequency band equals the estimated binaural masking level difference (BMLD) for a "comparable" tone-in-noise detection task. Specifically, if the difference in the IPD of T and M at the center frequency of band i is equal to x rad, the binaural advantage in band i is estimated as the expected BMLD when detecting a tone at the band center frequency in the presence of a diotic masker when the tone has an IPD of x rad. The maximum binaural advantage in a band [taken directly from Zurek, 1993, Fig. 15.2, and shown in Fig. 5(a) as a function of frequency] occurs when, at the band center frequency, the IPD of T and M differ by π rad. When the difference in the T and M IPD at the band center frequency is less than π rad, the binaural advantage in the band is lower (in accord with the Colburn model). The amount of information (γ_i) in each band (the "band efficiency") is computed as

$$\gamma_i = \begin{cases} 0, & R_i < -12 \text{ dB} \\ R_i + 12, & -12 \text{ dB} < R_i < 18 \text{ dB} \\ 30, & R_i > 18 \text{ dB} \end{cases} \quad (1)$$

This operation assumes that there is no incremental improvement in target audibility with increases in TMR above some asymptote (i.e., 18 dB) and no decrease in target audibility with additional decrements in TMR once the target is below masked threshold (i.e., −12 dB). The analysis implicitly assumes that the target is well above absolute threshold. Finally, the values of γ_i are multiplied by the frequency-dependent weights shown in Fig. 5(b) (which represent the relative importance of each frequency band for understanding speech) and summed to estimate the effective AI. The effective AI can take on values between 0.0 (if all R_i are less than or equal to 12 dB) and 1.0 (if all R_i are greater than or

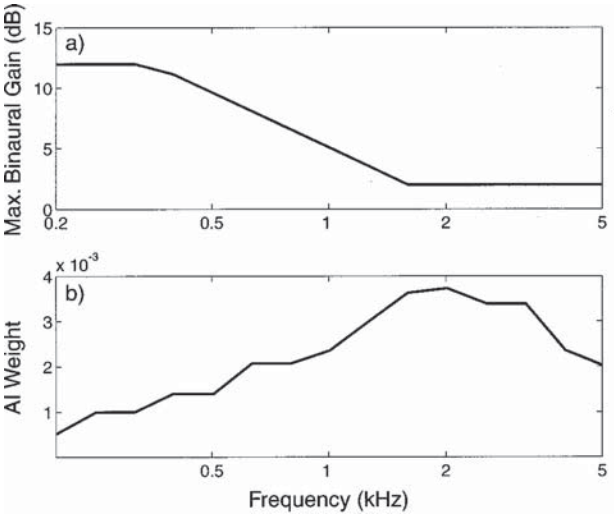

FIG. 5. Binaural AI model assumptions (Zurek, 1993). Panel (a) shows maximal binaural advantage (improvement in effective target-to-masker level ratio or TMR) as a function of frequency, which only arises when IPD of T and M differ by 180°. Panel (b) shows weighting of information at each frequency for speech intelligibility.

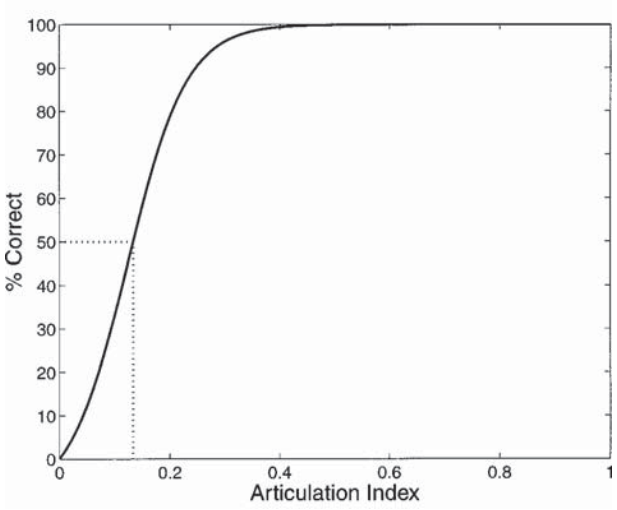

FIG. 6. Assumed relationship between AI and percent words correct assumed for high-context speech (as described in Hawley, 2000). Dashed lines show threshold level for the experiments reported herein.

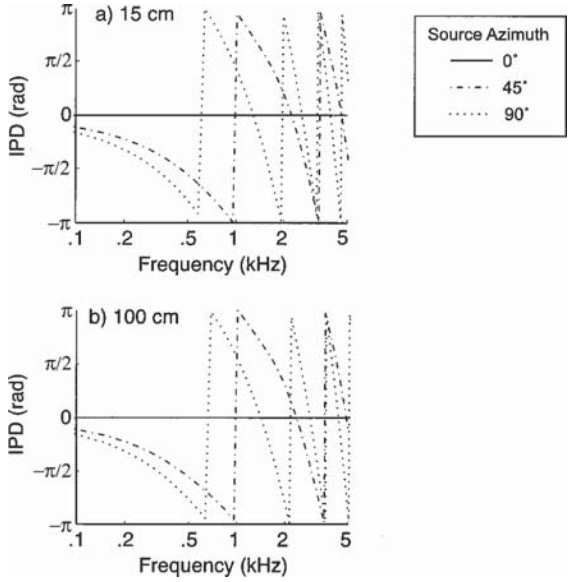

FIG. 7. Interaural phase differences as a function of frequency for the spherical-head HRTFs. (a) Near distance (15 cm) in top panel. (b) Far distance (1 m).

equal to 18 dB). For a given speech intelligibility task and a given set of speech materials, percent correct is a monotonic function of AI (e.g., see Kryter, 1962); for the high-context speech materials used in the present study, this correspondence, as derived by Hawley (2000), is shown in Fig. 6.

Using this model, Zurek (1993) was able to predict the spatial unmasking effects observed in a number of studies that used steady-state maskers (such as broadband noise) and positioned both T and M at a distance of at least 1 m from the subject (e.g., Dirks and Wilson, 1969; Plomp and Mimpen, 1981; Bronkhorst and Plomp, 1988, among others). In this paper, we apply this model to cases when the target and/or masker are close to the subject (i.e., 15 cm).

B. Predicted speech intelligibility at speech reception threshold

In order to calculate model predictions of the current results, the IPDs in the spherical-head HRTFs were analyzed. Figure 7, which plots the IPD in the HRTFs (as a function of frequency) for the positions used in the study, shows that IPD varies dramatically with source laterality and only slightly with distance (e.g., see Brungart and Rabinowitz, 1999; Shinn-Cunningham *et al.*, 2000). Using the left- and right-ear TMRs at the measured SRT (Fig. 3), the difference in T and M IPD was used to compute the effective TMR (the TMR at the better ear, adjusted for binaural gain) and the "band efficiency" in each frequency band. From these values, the AI was calculated and used to predict percentage correct key words using the mapping shown in Fig. 6.

We applied a similar analysis to the left and right ear stimuli in isolation (i.e., for a comparable configuration but with one of the ears "turned off"). To generate these monaural predictions, the appropriate monaural TMR (Fig. 3) was used to compute the AI directly (excluding any binaural contributions). In this way, we predicted not only the percentage-correct words for binaural stimuli but also left- and right-ear monaural stimuli.

Figure 8 shows the predicted percentage correct on our high-context speech task when the T and M levels equaled those presented at SRT. Predictions are shown for binaural listeners (x's) as well as monaural-left and monaural-right listeners (triangles and circles, respectively). The relative levels of T and M used in the predictions are those at which subjects correctly identified approximately 50% of the sentence key words. Thus, the model correctly predicts an observed result when the prediction is close to 50%. For our purposes, predictions falling within the gray area in each panel (within 10% of the defined 50%-correct threshold) are considered to match measured performance.[3] Note that in the

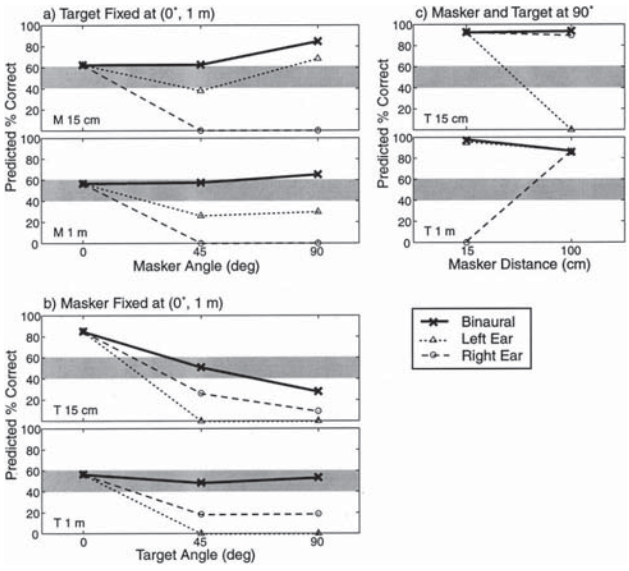

FIG. 8. Predicted percent-correct word scores from model using TMRs and binaural cues present at threshold (actual performance indicated by gray region). Bold exes show binaural model predictions; triangles and circles give monaural, left- and right-ear predictions, respectively. Conditions: (a) T fixed (0°, 1 m) and M at each of 6 locations; (b) M fixed (0°, 1 m) and T at each of 6 locations; and (c) T and M at 90° and 15 cm or 1 m.

model, predicted monaural performance (triangles or circles) is always less than or equal to binaural performance (exes), because any binaural processing will only increase the AI calculated from the better ear (and hence the predicted level of performance).

The one constant feature in Fig. 8 concerns the worse-ear monaural predictions. In every configuration for which the TMR differs in the two ears [four in Fig. 8(a) (circles), four in Fig. 8(b) (triangles), and two in Fig. 8(c) (rightmost triangle in top panel, leftmost circle in bottom panel)] the worse-ear, predicted percent correct is 0%.

Figure 8(a) shows predictions for T fixed ahead. For the diotic configurations [left side of Fig. 8(a)] both ears receive the same stimulus, left- and right-ear monaural predictions are identical, and there is no predicted benefit from listening binaurally. For all configurations in which M is at 1 m [lower panel, Fig. 8(a)], binaural predictions fall within or slightly above the expected range. Predictions for the better (left) ear are near 30% correct when the 1-m M is positioned laterally. When M is at 15 cm [upper panel in Fig. 8(a)], the binaural model predictions are generally higher than observed performance, but the error is only significant when M is at (90°, 15 cm) (binaural prediction near 90% correct). The monaural better-ear prediction is slightly below measured performance when M is at (45°, 15 cm) and substantially above measured performance when M is at (90°, 15 cm).

Figure 8(b) shows the predictions when M is fixed at (0°, 1 m). For this condition, the binaural predictions fit the data well for all configurations in which T is at the farther (1 m) distance [lower panel in Fig. 8(b)]. For the distant, laterally displaced T, better-ear predictions fall well below true binaural performance (19% correct for T at 45° and 90°). When T is at 15 cm, the binaural model predictions are less accurate, overestimating performance for T at 0° and underestimating performance for T at 90°.

In all four configurations in which T and M are positioned at 90° [Fig. 8(c)], the model predicts that both binaural performance and monaural better-ear performance should be much better than what was actually observed, with the predictions ranging from 86% to 95% correct.

C. Predicted spatial unmasking

The Zurek model (1993) was also used to predict the magnitude of the spatial unmasking in the various spatial configurations. To make these predictions, the mapping in Fig. 6 was used to predict the AI at which 50% of the key words are identified (see the dashed lines in Fig. 6). We then computed the level that T would have to emit in order to yield this threshold AI for each spatial configuration (assuming that the level emitted by M is fixed) and subtracted the level T would have to emit in the diotic reference configuration. Similar analysis was performed for left- and right-ear monaural signals in order to predict the impact of having only one functional ear.

Results of these predictions are shown in Fig. 9. In the figure, the large symbols show the mean unmasking found in the binaural experiments (presented previously in Fig. 4), while the lines with small symbols show the corresponding binaural (solid lines), left-ear (dashed lines), and right-ear

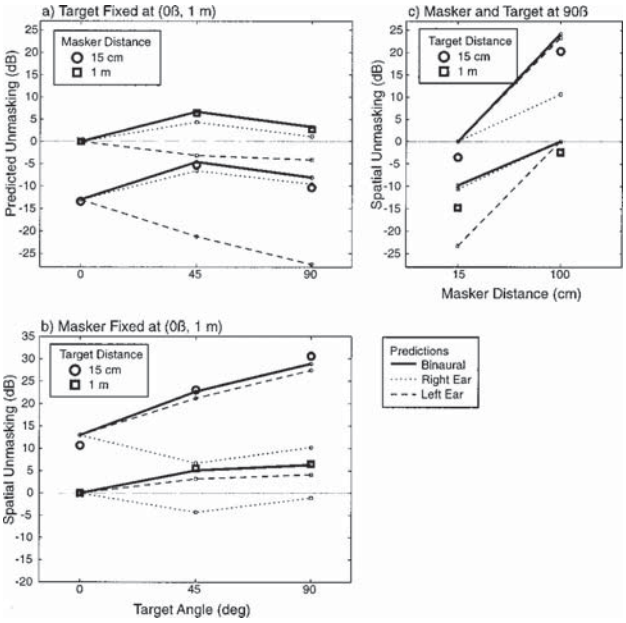

FIG. 9. Spatial advantage (energy a target emits at threshold for a constant-energy masker) and model predictions, relative to diotic reference. Symbols show across-subject means of measured spatial advantage, repeated from Fig. 4. Lines give model predictions: solid line for binaural model; dotted and dashed lines for left and right ears (without binaural processing), respectively. In any one configuration, the difference between the solid line and the better of the dotted or dashed lines gives the predicted binaural contribution to unmasking; the difference between the dotted and dashed lines yields the predicted better-ear advantage.

(dotted lines) predictions. To the extent that the model is accurate, the difference in binaural and better-ear predictions at each spatial configuration gives an estimate of the binaural contribution to spatial unmasking; the difference between the binaural and worse-ear predictions predicts how large the impact of listening with only one ear can be (i.e., if the acoustically better ear is nonfunctional).

The binaural predictions capture the main trends in the data, accounting for 99.05% of the variance in the measurements. The only binaural predictions that are not within the approximate 1-dB standard error in the measurements correspond to the same configurations for which the predicted percent-correct scores fail.

D. Difference between better- and worse-ear thresholds

The spatial unmasking analysis presented in Fig. 9 separately estimates binaural, monaural better-ear, and monaural worse-ear thresholds (in dB). From these values, we can predict the binaural advantage (i.e., the difference between the binaural and the better-ear threshold) and the difference between the better- and worse-ear thresholds (at least to the extend that the Zurek, 1993 model is accurate). These values are presented in Table I. The difference between the better- and worse-ear thresholds (second data column) is calculated as the absolute value of the difference (in dB) of the threshold T levels for left- and right-ear monaural predictions. This difference ranges from 5–18 dB for configurations in which T and M are not in the same location. Comparing these estimates (which weigh the TMR at each frequency according

to the AI calculation) to estimates made from the strict acoustic analysis (which weigh all frequencies up to 8000 Hz equally; first data column) shows (not unexpectedly) that the two methods yield very similar results. The predicted binaural advantage (third data column in Table I), defined as the difference between binaural and monaural better-ear model predictions for each configuration, is uniformly small, ranging from 0–2 dB.

E. Discussion

The Zurek model (1993) does a very good job of predicting the results for all spatial configurations similar to those that have been tested previously. In fact, the model fails only when T and/or M are near the head or when both T and M are located laterally.

Of the 15 independent spatial configurations tested, predicted performance is better than observed for six configurations, worse than observed for one configuration, and in agreement with the measurements in the remaining eight configurations. In six of the seven configurations for which the model prediction differs substantially from observed performance, T and/or M have ILDs that are larger than in previously tested configurations.

The Zurek model uses a simplified version of Colburn's model (1977a, b) of binaural unmasking to predict the binaural gain in each frequency channel, given the interaural differences in T and M. Colburn's original model accounts for the fact that binaural unmasking decreases with the magnitude of the ILD in M because the number of neurons contributing binaural information decreases with increasing ILD. The simplified version of the Colburn model used in Zurek's formulation does not take into account how the noise ILD affects binaural unmasking. If one were to use a more complex version of the Colburn binaural unmasking model, the predicted binaural gain would be smaller for spatial configurations in which there is a large ILD in the masker. Binaural predictions from such a corrected model would fall somewhere between the current binaural and better-ear predictions.

Unfortunately, such a correction will not improve the predictions. In particular, of the seven predictions that differ substantially from the measurements, there is only one case in which decreasing the binaural gain in the model prediction could substantially improve the model fit [T at (0°, 1 m) and M at (90°, 15 cm); see Fig. 9(a), circle at right side of panel]. In five of the remaining configurations in which the predictions fail [circle symbol at left of Fig. 9(b) and all four observations in Fig. 9(c)], even the better-ear model analysis predicts more spatial unmasking than is observed, and in the final configuration [e.g., circle symbol at right of Fig. 9(b)] both the binaural and better-ear analysis predict less unmasking than was observed. In fact, for this configuration, any decrement in the binaural contribution of the model will degrade rather than improve the binaural prediction fit.

The model assumes that binaural processing can only improve performance above what would be achieved if listening with the better ear alone. Current results suggest that this may not always be the case; we found that measured binaural performance is sometimes worse than the predicted performance using the better ear alone. We know of only one study that found a binaural *dis*-advantage for speech unmasking. Bronkhorst and Plomp (1988) manipulated the overall interaural level differences of the signals presented to the subjects in order to simulate monaural hearing loss. Subjects were tested with binaural, better-ear monaural, and worse-ear monaural stimuli as well as conditions in which the total signal to one of the ears was attenuated by 20 dB. In some cases, monaural performance using only the better-ear stimulus was near binaural performance; in these cases, attenuating the worse ear stimulus by 20 dB had a negligible impact on performance. If both ears had roughly the same TMR but the IPDs in T and M differed, binaural performance was best, performance for left- and right-ear monaural conditions was equal (and worse than binaural performance), and attenuating either ear's total stimulus caused a small (1–2 dB) degradation in SRT. Of most interest, in conditions for which there was a clear ''better ear'' (i.e., when the TMR was much larger in one ear than the other), performance with the better ear attenuated by 20 dB was worse than monaural performance for the better-ear stimulus, even though the better-ear stimulus was always audible. The researchers noted that this degradation in performance appears to be ''due to a ''disturbing'' effect of the relatively loud noise presented in the other ear'' (Bronkhorst and Plomp, 1988, p. 1514), because the better-ear stimulus played alone yielded better performance than the binaural stimulus. In the current experiment, some of the configurations for which the binaural predictions exceeded observed performance had a worse-ear signal that was substantially louder than the better-ear signal. However, when T was at (90°, 15 cm) and M was at (90°, 1 m), binaural performance was worse than predicted better-ear performance, even though the worse-ear signal was quieter than the better-ear signal. One possible explanation for these results is that large ILDs in the stimuli can sometimes degrade binaural performance below better-ear monaural performance, even if the worse-ear stimulus is quieter than the better-ear stimulus.

Finally, it should be pointed out that while the overall rms level of the stimuli was held constant at the better ear, the spectral content in T and M changed with spatial position as a result of the HRTF processing. It may be that some of the prediction errors arise from problems with the monaural, not binaural, processing in the model. Further experiments are needed to directly test whether binaural performance is worse than monaural better-ear performance in spatial configurations like those tested.

V. CONCLUSIONS

The results of these experiments demonstrate that the amount of spatial unmasking that can arise when T and/or M are within 1 m of a listener is dramatic. For a masker emitting a fixed-level noise, the level at which a speech target must be played to reach the same intelligibility varies over approximately 45 dB for the spatial configurations considered. Much of this effect is the result of simple changes in stimulus level with changes in source distance; however, other phenomena also influence these results.

It is well known that, on spatial unmasking tasks, monaural listeners are at a disadvantage compared to binaural listeners. In roughly half of the possible spatial configurations, the better-ear advantage is lost and any binaural processing gains are ineffective for these listeners (e.g., see Zurek, 1993). However, the current results suggest that when either T or M are close to the listener, monaural listeners can suffer from disadvantages (compared to normal-hearing listeners) that are as much as 13 dB greater observed for configurations in which T and M are at least 1 meter from the listener [i.e., from Table I, when T is at (0°, 1 m), the estimated left/right asymmetry is 19.6 dB for M at (90°, 15 cm) and only 6.4 for M at (90°, 1 m)]. Specifically, for the configurations tested, the worse-ear TMR can be nearly 20 dB lower than the better-ear TMR. While the current experiments did not measure performance of monaural listeners directly, this analysis supports the view that having two ears provides an enormous advantage to listeners in noisy environments, especially when the sources of interest are close to the listener. However, much of the benefit obtained from listening with two ears appears to derive from having two independent ''mixes'' of T and M, one of which often has a better TMR than the other. The specifically binaural processing advantages expected in the tested configurations are comparable to those observed in previous studies, on the order of 2 dB. Of course, even 2 dB of improvement in TMR can lead to vast improvements in speech intelligibility near SRT, leading to improvements in percent-correct word identification of over 20%.

The current experiments included a number of novel spatial configurations that have not previously been investigated. For many of these configurations, the Zurek model of spatial unmasking of speech fails to predict observed performance. The reasons underlying these failures (which all simulate either T or M very near the listener or have both T and M located at 90°) must be investigated further. One of the failed predictions may be partially corrected by considering a binaural unmasking model that takes into account the ILD in the masker [i.e., when M is at (90°, 15 cm) and T is at (0°, 1 m)]. However, such a correction will not improve the model predictions for any of the remaining configurations for which the model fails.

Analysis suggests that binaural processing of interaural phase decreases SRT by 1–2 dB for the configurations considered in the current study, similar to the gain observed for configurations in which T and M are both at least 1 meter from the listener (e.g., see Bronkhorst, 2000). However, for the configurations in which better-ear monaural predictions of SRT are lower than the SRTs observed with binaural presentations, there may actually be a disadvantage to listening with two ears (compared to listening with the better ear alone). Additional experiments using monaural control conditions must be performed in order to fully explore whether large ILDs degrade speech intelligibility or whether monaural better-ear performance is worse than predicted in these configurations.

ACKNOWLEDGMENTS

This work was supported in part by AFOSR Grant No. F49620-98-1-0108 to B.G.S.C. and NIDCD Grant No. DC02696 to R.Y.L. Portions of this work were presented at the Spring 2000 meeting of the Acoustical Society of America. Les Bernstein, Bob Gilkey, and an anonymous reviewer provided very helpful, constructive comments on earlier versions of this manuscript.

[1]Intersubject differences were relatively modest in these experiments, with an average sample standard deviation across the four subjects of 1.7 dB. These subject differences are shown in Fig. 4, but are left off of Figs. 3 and 4 for clarity.

[2]The AI has since been extended and renamed as the Speech Intelligibility Index or SII; see ANSI, 1997.

[3]On average, the standard error in the SRTs across the four subjects is 0.85 dB. From this, we can estimate the corresponding standard error in percent-correct estimates as follows. Near the 50%-correct point, the AI function is roughly linear. Assuming all frequency bands have TMRs between −12 and 18 dB, the AI is linear with TMR. Thus, under these assumptions (i.e., near the 50%-correct point with all frequency bands contributing to the AI and not saturated), percent correct is a linear function of TMR with a slope of roughly 12%/dB. Multiplying the standard error in the estimate of SRT times this slope yields a rough approximation of the standard error in the percentage correct of 0.85*12%=10.2%. Note that if some frequency bands are inaudible or saturated, the estimated error in percent correct will actually be less than 10.2%.

ANSI (**1969**). ANSI 53.5-1969, ''American National Standard Methods for the Calculation of the Articulation Index'' (American National Standards Institute, New York).

ANSI (**1997**). ANSI 53.5-1997, ''Methods for Calculation of the Speech Intelligibility Index'' (American National Standards Institute, New York).

Bronkhorst, A. W. (**2000**). ''The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,'' Acustica **86**, 117–128.

Bronkhorst, A. W., and Plomp, R. (**1988**). ''The effect of head-induced interaural time and level differences on speech intelligibility in noise,'' J. Acoust. Soc. Am. **83**, 1508–1516.

Bronkhorst, A. W., and Plomp, R. (**1990**). ''A clinical test for the assessment of binaural speech perception in noise,'' Audiology **29**, 275–285.

Brown, T. J. (**2000**). ''Characterization of Acoustic Head-Related Transfer Functions for Nearby Sources,'' M.Eng. thesis, Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Brungart, D. S., and Rabinowitz, W. M. (**1999**). ''Auditory localization of nearby sources. I. Head-related transfer functions,'' J. Acoust. Soc. Am. **106**, 1465–1479.

Colburn, H. S. (**1977a**). ''Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise,'' J. Acoust. Soc. Am. **64**, 525–533.

Colburn, H. S. (**1977b**). ''Theory of binaural interaction based on auditory-nerve data. II. Detection of tones in noise,'' J. Acoust. Soc. Am. **61**, 525–533. See Aip Document No. E-PAPS JASMA-6-525-98.

Dirks, D. D., and Wilson, R. H. (**1969**). ''The effect of spatially separated sound sources on speech intelligibility,'' J. Speech Hear. Res. **12**, 5–38.

Duda, R. O., and Martens, W. L. (**1997**). ''Range-dependence of the HRTF for a spherical head,'' IEEE ASSP Workshop on Applications of DSP to Audio and Acoustics.

Durlach, N. I., and Colburn, H. S. (**1978**). ''Binaural phenomena,'' in *Handbook of Perception*, edited by E. C. Carterette and M. P. Friedman (Academic, New York), Vol. 4, pp. 365–466.

Fletcher, H., and Galt, R. H. (**1950**). ''The perception of speech and its relation to telephony,'' J. Acoust. Soc. Am. **22**, 89–151.

Good, M. D., Gilkey, R. H., and Ball, J. M. (**1997**). ''The relation between detection in noise and localization in noise in the free field,'' in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. Gilkey and T. Anderson (Erlbaum, New York), pp. 349–376.

Hawley, M. L. (**2000**). ''Speech Intelligibility, Localization and Binaural Hearing in Listeners with Normal and Impaired Hearing,'' Ph.D. dissertation, Biomedical Engineering, Boston University.

Hawley, M. L., Litovsky, R. Y., and Colburn, H. S. (**1999**). ''Speech intelligibility and localization in a multi-source environment,'' J. Acoust. Soc. Am. **105**, 3436–3448.

Hirsh, I. J. (**1950**). ''The relation between localization and intelligibility,'' J. Acoust. Soc. Am. **22**, 196–200.

IEEE (**1969**). ''IEEE recommended practice for speech quality measurements,'' IEEE Trans. Audio Electroacoust. **17**(3), 225–246.

Kryter, K. D. (**1962**). ''Methods for the calculations and use of the Articulation Index,'' J. Acoust. Soc. Am. **34**, 1689–1697.

MacKeith, N. W., and Coles, R. R. A. (**1971**). ''Binaural advantages in hearing of speech,'' J. Laryngol. Otol. **85**, 213–232.

Pavlovic, C. V. (**1987**). ''Derivation of primary parameters and procedures for use in speech intelligibility predictions,'' J. Acoust. Soc. Am. **82**, 413–422.

Peissig, J., and Kollmeier, B. (**1997**). ''Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners,'' J. Acoust. Soc. Am. **101**, 1660–1670.

Plomp, R., and Mimpen, A. M. (**1981**). ''Effect of the orientation of the speaker's head and the azimuth on a noise source on the speech reception thresholds for sentences,'' Acustica **48**, 325–328.

Rabinowitz, W. R., Maxwell, J. Shao, Y., and Wei, M. (**1993**). ''Sound localization cues for a magnified head: Implications from sound diffraction about a rigid sphere,'' Presence **2**(2), 125–129.

Shinn-Cunningham, B. G. (**2000**). ''Distance cues for virtual auditory space,'' in Proceedings of the IEEE-PCM 2000, pp. 227–230, Sydney, Australia, 13–15 December.

Shinn-Cunningham, B. G., Santarelli, S., and Kopčo, N. (**2000**). ''Tori of confusion: Binaural localization cues for sources within reach of a listener,'' J. Acoust. Soc. Am. **107**, 1627–1636.

Zurek, P. M. (**1993**). ''Binaural advantages and directional effects in speech intelligibility,'' in *Acoustical Factors Affecting Hearing Aid Performance*, edited by G. Studebaker and I. Hochberg (College-Hill, Boston).

Object continuity enhances selective auditory attention

Virginia Best, Erol J. Ozmeral, Norbert Kopčo, and Barbara G. Shinn-Cunningham*

Department of Cognitive and Neural Systems, Boston University, 677 Beacon Street, Boston, MA 02215

Edited by Eric I. Knudsen, Stanford University School of Medicine, Stanford, CA, and approved June 27, 2008 (received for review April 16, 2008)

In complex scenes, the identity of an auditory object can build up across seconds. Given that attention operates on perceptual objects, this perceptual buildup may alter the efficacy of selective auditory attention over time. Here, we measured identification of a sequence of spoken target digits presented with distracter digits from other directions to investigate the dynamics of selective attention. Performance was better when the target location was fixed rather than changing between digits, even when listeners were cued as much as 1 s in advance about the position of each subsequent digit. Spatial continuity not only avoided well known costs associated with switching the focus of spatial attention, but also produced refinements in the spatial selectivity of attention across time. Continuity of target voice further enhanced this buildup of selective attention. Results suggest that when attention is sustained on one auditory object within a complex scene, attentional selectivity improves over time. Similar effects may come into play when attention is sustained on an object in a complex visual scene, especially in cases where visual object formation requires sustained attention.

source segregation | auditory scene analysis | spatial hearing | streaming | auditory mixture

In everyday situations, we are confronted with multiple objects that compete for our attention. Both stimulus-driven and goal-related mechanisms mediate the between-object competition to determine what will be brought to the perceptual foreground (1, 2). In natural scenes, objects come and go and the object of interest can change from moment to moment, such as when the flow of conversation shifts from one talker to another at a party. Thus, our ability to analyze objects in everyday settings is directly affected by how switching attention between objects affects perception. Much of what we know about the effects of switching attention comes from visual experiments in which observers monitor rapid sequences of images or search for an item in a static field of objects (3, 4). Although these situations give insight into the time it takes to dis- and reengage attention from one object to the next, they do not directly explore whether there are dynamic effects of sustaining attention on one object through time.

In contrast to visual objects, the identity of an auditory object is intimately linked to how the content of a sound evolves over time. Moreover, the process of forming an auditory object is known to evolve over seconds (5–8). Given that attention is object-based (9, 10), this refinement in object formation may directly impact the selectivity of attention in a complex auditory scene. Specifically, sustaining attention on one object in a complex scene may yield more refined selectivity to the attended object over time. In turn, switching attention to a new object may reset object formation and therefore reset attentional selectivity. If so, the cost of switching attention between objects may not only be related to the time required to dis- and reengage attention (3, 11, 12) but also to the time it takes to build up an estimate of the identity of an object in a scene.

In the current study, we measured how switching spatially directed attention influenced the ability to recall a sequence of spoken digits. Five loudspeakers were distributed horizontally in front of the listener. Listeners identified sequences of four digits presented either from one loudspeaker or from a different loudspeaker chosen randomly on each digit, with visual cues indicating the target loudspeaker at each temporal position in the sequence. The remaining four loudspeakers presented simultaneous distracter digits. To explore whether continuity of a nonspatial feature influenced performance, we tested conditions in which the target voice changed from digit to digit (Exp. 1) as well as conditions under which the target voice was the same from digit to digit (Exp. 2). We investigated the time course of the cost of switching attention by testing four different overall rates of presentation, obtained by varying the silent delays inserted between each digit in the sequence (0, 250, 500, or 1,000 ms). To determine whether advance knowledge of where to redirect spatial attention ameliorated some of the cost of switching attention, we compared conditions under which the visual indicator of target location was turned on synchronously with the digits to those in which the visual cue preceded the auditory stimuli by the corresponding interdigit delay.

Results suggest that sustaining attention on one continuous auditory stream leads to refinements in selective attention over time. This refinement in selective attention is lost when attention switches to a new object, adding to the cost of switching attention between objects in a complex scene.

Results

In both experiments at all interdigit delays, mean performance was better when the spatial location of the target did not change between digits (the "fixed" condition, F) than when listeners had to instantaneously switch attention to a new location for each digit (the "switching, LED synchronous" or SS condition) (Fig. 1, compare squares and circles). Moreover, performance in the SS condition tended to be better at slower presentation rates than at faster rates, when there was time to dis- and reengage spatially directed attention to the new digit position. The cost of switching spatial attention to a new location was thus positive in both experiments for all presentation rates and decreased with decreasing presentation rate (Fig. 2, circles). However, even at the slowest presentation rate, when there was 1 s of silence between subsequent digits, a switching cost was evident. In general, continuity of voice across digits (Exp. 2) (Figs. 1 *Lower* and 2 *Lower*) increased the cost of switching spatial attention compared with when voice quality changed between target digits (Exp. 1) (Figs. 1 *Upper* and 2 *Upper*). This improvement with voice continuity was especially pronounced at the shortest interdigit delays, where the temporal continuity between the target digits was greatest.

Author contributions: V.B., E.J.O., N.K., and B.G.S.-C. designed research; V.B., E.J.O., and N.K. performed research; V.B. and E.J.O. analyzed data; and V.B. and B.G.S.-C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

*To whom correspondence should be addressed. E-mail: shinn@cns.bu.edu.

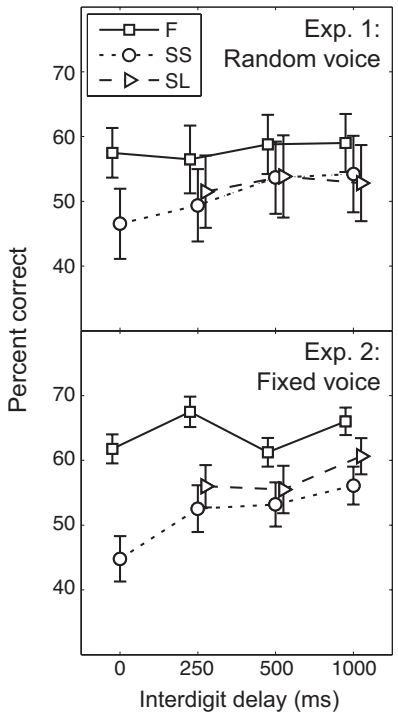

Fig. 1. Overall performance is best when spatial location is fixed between digits; moreover, even up to 1 s of advance knowledge of where to direct spatial attention does not overcome the cost of switching spatial attention. Across-subject mean scores (±SEM) for Exp. 1, where the target voice switches between digits (*Upper*), and Exp. 2, where the target voice is fixed across digits (*Lower*). Data are plotted as a function of interdigit delay for conditions F (squares and solid lines), SS (circles and dotted lines), and SL (triangles and dashed lines).

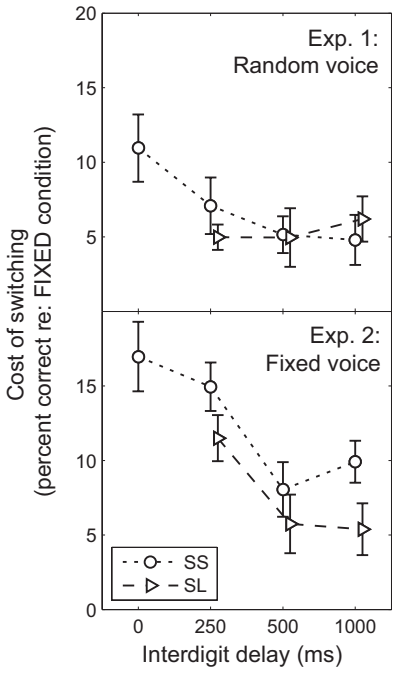

Fig. 2. The cost of switching spatial attention decreases with interdigit delay but is always positive. Moreover, the cost of switching tends to be greater when voice quality is fixed between digits (Exp. 2) (*Lower*) than when the voice changes between digits (Exp. 1) (*Upper*), especially at short interdigit delays. Each plot shows the across-subject mean difference in performance (±SEM) between condition F and each of the conditions SS (circles and dotted lines) and SL (triangles and dashed lines).

We predicted that providing spatial information in advance during the gaps between digits in the target sequence would eliminate the cost of switching spatial attention. In the "switching, LED leading (SL)" condition, the LEDs were turned on at the beginning of the silent gap preceding a target digit (see *Materials and Methods*). Surprisingly, when the target voice switched between target digits (Exp. 1), there was no reduction in the cost of switching spatial attention with advance warning about where the next target digit would be (Figs. 1 *Upper* and 2 *Upper*, compare circles and triangles). In contrast, when the target voice was fixed throughout a trial (Exp. 2), the cost of switching spatial attention was reduced, but not eliminated, by advance knowledge of target location (Figs. 1 *Lower* and 2 *Lower*, compare circles and triangles).

An examination of performance as a function of temporal position within the four-digit sequence revealed that the cost associated with switching the target location was not constant across time (Fig. 3). For the switching conditions, performance tended to be better for the first and last digit (see roughly U-shaped functions in Fig. 3, circles and triangles), consistent with typical primacy/recency effects on memory tasks. In contrast, for the F condition, the first digit was identified the most poorly and the remaining three digits were identified with increasing accuracy (Fig. 3, squares). In other words, the cost of switching spatially directed attention tended to increase throughout the duration of the sequence. This was particularly true for the faster rates when the target voice was held constant (Fig. 3 *Lower*, two left plots).

Statistical comparison of performance in the F and SS conditions revealed significant main effects of condition [$F(1, 4) = 19.6, P < 0.05$], delay [$F(3, 12) = 20.9, P < 0.001$], and temporal position [$F(3, 12) = 7.9, P < 0.005$], as well as significant two-way interactions between condition and delay [$F(3, 12) = 7.0, P < 0.01$], condition and temporal position [$F(3, 12) = 11.8, P < 0.05$], and delay and temporal position [$F(9, 36) = 2.4, P < 0.05$] in Exp. 1. In Exp. 2, significant main effects of condition [$F(1, 4) = 55.8, P < 0.005$], delay [$F(3, 12) = 22.4, P < 0.001$], and temporal position [$F(3, 12) = 10.7, P < 0.005$] were found. All two-way interactions were also significant [condition and delay: $F(3, 12) = 38.0, P < 0.001$; condition and temporal position: $F(3, 12) = 40.3, P < 0.001$; delay and temporal position: $F(9, 36) = 3.7, P < 0.005$], as was the three-way interaction [$F(9, 36) = 5.9, P < 0.001$].

The influence of the preceding visual cue (compare circles and triangles in Fig. 3) was negligible for all temporal positions in Exp. 1 but led to improved performance in Exp. 2 for later temporal positions and longer delays. This was supported by statistical comparison of performance under the SS and SL conditions, which found a significant main effect of delay in Exp. 1 [$F(2, 8) = 6.4, P < 0.05$] but no other significant effects or interactions, and significant main effects of condition [$F(1, 4) = 42.7, P < 0.005$] and delay [$F(2, 8) = 16.5, P < 0.005$] in Exp. 2, as well as significant two-way interactions between condition and temporal position [$F(3, 12) = 6.7, P < 0.01$] and delay and temporal position [$F(6, 24) = 3.8, P < 0.01$], and a significant three-way interaction [$F(6, 24) = 2.8, P < 0.05$].

An analysis of incorrect responses revealed that subjects had a tendency to report digits that were presented from loudspeakers adjacent to the target loudspeaker when they did not correctly identify the target (Fig. 4 *Upper*). Responses to masker digits decreased as the distance between the masker loudspeaker and the cued, target loudspeaker increased. The number of responses that did not correspond to either the target digit or one of the simultaneous masker digits was relatively low ("rand"; note that if subjects randomly guessed among all possible

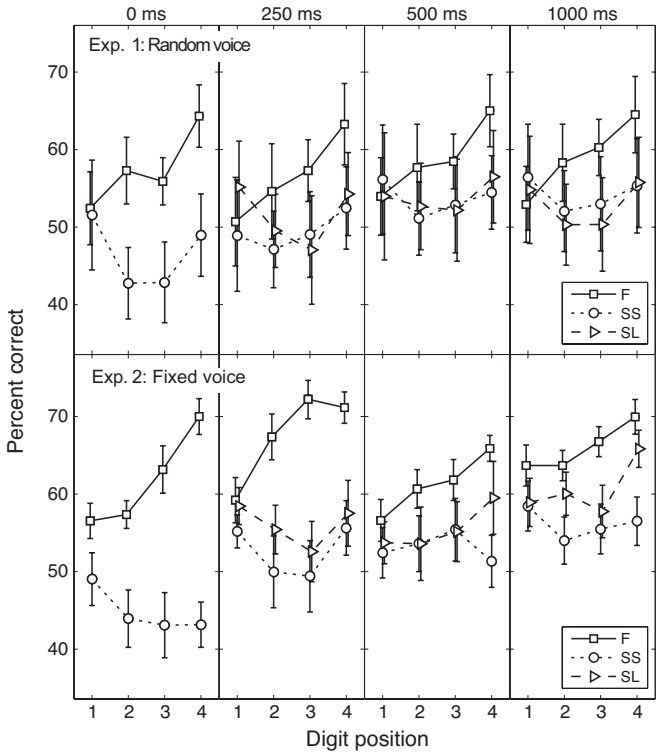

Fig. 3. When the target sequence is continuous in spatial location, performance improves from digit to digit, an effect that is enhanced when the target voice quality is continuous between digits. Across-subject mean scores (±SEM) as a function of temporal position for Exp. 1 (with random voice) (*Upper*) and Exp. 2 (with fixed voice) (*Lower*). The four plots within each row show data for the four different interdigit delays. Data are plotted as a function of temporal position within the target sequence for F (squares and solid lines), SS (circles and dotted lines), and SL (triangles and dashed lines).

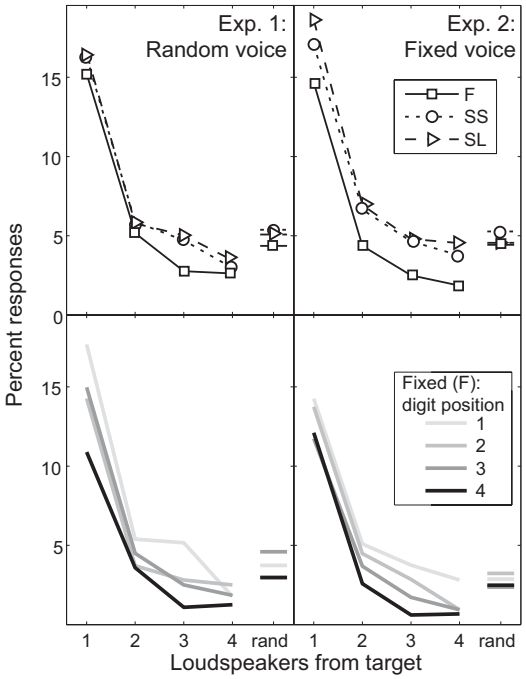

Fig. 4. Spatially directed attention filters out sources from the wrong direction, and this filtering becomes more refined over time when target location is fixed across digits. (*Upper*) Percentage of responses that corresponded to a digit presented from a nontarget loudspeaker are shown as a function of the distance between the target loudspeaker and the loudspeaker presenting the reported digit. Responses that did not correspond to any of the presented digits are shown at the far right (rand). Responses are pooled across all subjects and all delays for F (squares and solid lines), SS (circles and dotted lines), and SL (triangles and dashed lines). (*Lower*) Incorrect responses in the F condition as a function of distance between the target loudspeaker and the loudspeaker presenting the reported digit for each temporal position within the sequence (light to dark gray showing results for target digits 1–4). Responses are pooled across all subjects and all delays.

answers when they were unsure of the target digit, this kind of error would be the most common). In the F condition, the improvement in performance across time came about primarily from a decrease in responses to digits presented from masker loudspeakers (Fig. 4 *Lower*).

Discussion

When identifying speech in the presence of competitors, attention to features such as voice and location can guide selective attention (13–18). The current results demonstrate that continuity of these perceptual features, which help define an object's identity, lead to improvements over time in the ability to select a target sequence from a complex acoustic scene. We suggest that this improvement in selective attention occurs because attention operates on perceptual objects, and the identity of an acoustic object in a complex scene depends on evidence acquired over the course of several seconds. Of course, feature-based attention could also account for the basic pattern of our results, but only if listeners can direct attention to multiple features simultaneously.

Slowing the presentation rate of a sequence of target digits reduces some of the cost associated with switching, consistent with there being a finite time required to disengage and then reengage attention (19, 20). However, delays as long as 1 s did not eliminate the cost of switching attention, suggesting that this cost was not entirely due to the time required to redirect attention. Moreover, performance improved over time for a target with continuity of perceptual features; disrupting object continuity reset this across-time refinement. Spatial continuity can also enhance auditory selective attention over much longer time scales (21). These results suggest that listeners refine selective auditory attention over time in a complex acoustic mixture.

The pattern of errors observed in these experiments shows that listeners were particularly susceptible to reporting masker words that occurred simultaneously from locations adjacent to the target. This pattern of errors is consistent with a popular model of spatial attention in which attention is directed via a tuned filter having a spatial focus and some finite spatial extent (e.g., see refs. 22 and 23). For the task and conditions tested here, it appears that the spatial attentional filter is sufficiently broad that adjacent locations are imperfectly rejected. However, we also find that the spatial filter becomes more focused over time when the target location is fixed from digit to digit (see also ref. 24).

Comparison of results from Exps. 1 and 2 suggests that continuity of voice enhances the benefit of spatial continuity of the target sequence (i.e., the cost of switching is greater in Exp. 2 than in Exp. 1) (Fig. 2, compare *Upper* with *Lower*). This enhancement is greatest when interdigit delays are brief and the target digit sequence is relatively connected (continuous) across time. As noted above, feature-based attention could help explain these results; however, it is difficult to see how feature-based attention could account for this effect of stimulus timing. We find that any manipulation that enhances object formation causes an improvement in selective attention over time, whether it is continuity of a stimulus feature (spatial location, voice

quality) or a rapid presentation rate. Thus, parsimony favors the hypothesis that selective attention becomes increasingly more effective as object formation builds.

When the target sequence has spatial continuity and maximal voice continuity (Fig. 3 *Lower*, leftmost plot), performance for the first digit in the sequence is better than when spatial location changes between digits. This kind of effect can only be explained if the overall difficulty of a trial impacts how well the first digit of the target sequence is recalled at the conclusion of the trial, because the subject has no advance knowledge about the target location or target voice for the first digit in either the F or SS conditions. This result suggests that attentional demands are smallest when the target sequence is temporally connected, continuous in voice quality, and from a fixed location, leaving more resources for storage and recall of the sequence. This effect undoubtedly depends on overall memory demands of the task, and thus is likely to vary with the length of the target sequence as well as the listener's knowledge about when the sequence will end.

These findings shed light on why, in listening environments such as noisy parties or restaurants, it is more difficult to follow a conversation involving many people (where the relevant talker often and unexpectedly changes locations) than to focus on one talker (at one location) exclusively. In addition, these results may have implications for visual attention in tasks where object formation and target segmentation is challenging, or where the identity of a visual object depends on continuity of visual features over time (25).

Materials and Methods

Subjects. Five subjects (2 male, 3 female, aged 23–39 years) participated in Exp. 1. Five subjects (2 male, 3 female, aged 24–30 years) participated in Exp. 2, two of whom had participated in Exp. 1 before commencing Exp. 2 (S1 and S2). Subjects S1 and S2 were also two of the experimenters and had previously participated in several similar experiments. The other subjects were paid for their participation. All subjects were screened to ensure that they had normal hearing (within 10 dB) for frequencies between 250 Hz and 8 kHz. Experiments were approved by the Boston University Charles River Campus Institutional Review Board.

Environment. The experiments took place in a single-walled Industrial Acoustics Company booth with interior dimensions of 12′4″ × 13′ × 7′6″ (length × width × height), with perforated metal panels on the ceiling and walls and a carpeted floor (for an acoustic analysis of this environment, see ref. 26). The subject was seated on a chair in the center of the room. A head rest attached to the back of the chair cradled the neck and the back of the head to minimize head movements. No instructions were given to subjects regarding eye fixation during stimulus delivery, and eye movements were not measured. Stimuli were presented via five loudspeakers (215PS; Acoustic Research) located on a horizontal arc ≈5 ft from the subject at the level of the ears. The loudspeakers were positioned within the visual field of the subject, at lateral angles of −30°, −15°, 0°, 15°, and 30°. Subjects indicated their response by using a handheld keypad with an LCD display (QTERM). The booth was kept dark during the experiment, except for a small lamp placed on the floor behind the subject, which helped him or her to see the keypad.

Digital stimuli were generated and selected via a PC located outside the booth, and fed through five separate channels of Tucker-Davis Technologies hardware. Signals were converted at 20 kHz by a 16-bit D/A converter (DA8), attenuated (PA4), and passed through power amplifiers (Tascam) before presentation to the loudspeakers. Each loudspeaker had an LED affixed on its top surface, which could be turned on and off via the PC with a custom-built switchboard. MATLAB (Mathworks) software was used for stimulus generation, stimulus presentation, data acquisition, and analysis.

Stimuli. Stimuli consisted of the digits 1–9 spoken by 15 different male talkers from the TIDIGIT database (27). The mean duration of the set of digits was 434 ms (±103 ms). For each trial, five different sequences of four digits were presented simultaneously from the five spatially separated loudspeakers. For each of the four temporal positions in the sequence, the five digits were chosen randomly with the limitation that they were all different and spoken by a different talker. Digits were presented with synchronous onsets and were zero-padded at the end so that within each temporal position; all were the length of the longest digit in that particular position.

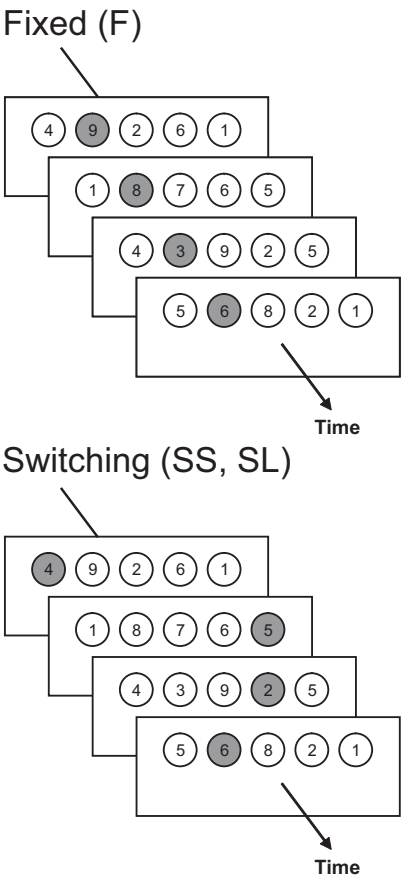

Fig. 5. Schematic of the auditory and visual stimuli for the fixed and switching conditions. Five different digits were presented simultaneously from the five loudspeakers (circles) in each of four temporal positions of the stimulus. During each of the four temporal positions, the LED on one loudspeaker was illuminated (filled circle) to indicate the target digit. (*Upper*) In the fixed condition, the target digit came from the same loudspeaker in each of the temporal positions. (*Lower*) In the switching conditions, the target came from a different random loudspeaker in each temporal position. The visual cue from the target LED came on simultaneously with the auditory stimuli in the F and SS conditions but preceded the auditory stimuli in the SL condition (diagram not shown).

One digit in each temporal position was designated as the target, with the four targets in the different temporal positions making up the target sequence. In each temporal position, one of the five LEDs was illuminated to indicate which loudspeaker contained the target. In the fixed condition (Fig. 5 *Upper*), this was the same loudspeaker for the whole sequence (although the loudspeaker varied randomly from trial to trial). In the two switching conditions (Fig. 5 *Lower*), the target loudspeaker was different in each temporal position so that the four digits in the sequence came from four different loudspeakers.

Conditions. In different experimental blocks, the sequences in a trial were presented with a different delay between the four digits (0, 250, 500, or 1,000 ms). This gave rise to average presentation rates of 2.3, 1.5, 1.1, and 0.7 words per second, respectively (although the variable digit lengths meant that the rhythm was not perfectly regular).

In the F and SS conditions, the LED turned on and off synchronously with the onset and offset of the digits in each temporal position. In the SL condition, the LED came on before the digits in each temporal position, with a lead time equal to the interdigit delay.

In Exp. 1, the voices were chosen randomly for each temporal position with the constraint that the same voice was not presented simultaneously from more than one loudspeaker. As a result, the target voice varied randomly throughout a target sequence. In Exp. 2, the four target digits in a sequence

were spoken by the same voice (chosen randomly on each trial). The maskers were chosen from the remaining 14 voices (separately for each temporal position).

Procedures. In an experimental test, the subject's task was to follow the LEDs and report the four-digit target sequence. Responses were entered by using the handheld keypad after the entire stimulus was finished. Subjects were forced to respond with a four-digit sequence and were instructed to guess the content for any digit that they did not hear. The sequence was scored on a per-digit basis in all analyses.

Each subject completed five sessions in an experiment, each on a separate day. A session consisted of one block of trials per combination of condition (F, SS, and SL) and delay (0, 250, 500, and 1,000 ms). Because the SS and SL conditions were identical for the 0-ms delay, there were 11 blocks of trials in total. The order of the blocks was random and different between sessions and subjects. A message on the keypad at the beginning of each block indicated which condition and delay would be presented in that block. Each block consisted of 40 trials.

Subjects did not complete any formal practice blocks, but were given exemplars of the stimuli to listen to while the experiment was being explained.

Statistical Analyses. The percentage correct data were arcsin transformed and submitted to two repeated measures ANOVAs. The first examined the effect of switching and had factors of condition (F and SS), interdigit delay (0, 250, 500, and 1,000 ms), and temporal position (1–4). The second compared performance in the two switching conditions using only the data that were independent (i.e., excluding the 0-ms delay).

ACKNOWLEDGMENTS. We thank Gerald Kidd and Chris Mason for many helpful discussions about both the experimental design and the data. This work was supported by Office of Naval Research Grant N000140710061 and National Institute on Deafness and Other Communication Disorders Grant R01 DC04663 (to B.G.S.-C.). V.B. was supported in part by a University of Sydney Postdoctoral Research Fellowship. N.K. was supported in part by Slovak Science Agency (VEGA) Grant 1/3134/06.

1. Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. *Annu Rev Neurosci* 18:193–222.
2. Yantis S (2005) How visual salience wins the battle for awareness. *Nat Neurosci* 8:975–977.
3. Kelley T, Serences J, Giesbrecht B, Yantis S (2008) Cortical mechanisms for shifting and holding visuospatial attention. *Cereb Cortex* 18:114–125.
4. Jefferies LN, Ghorashi S, Kawahara J, Lollo VD (2007) Ignorance is bliss: The role of observer expectation in dynamic spatial tuning of the attentional focus. *Percept Psychophys* 69:1162–1174.
5. Bregman AS (1990) *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).
6. Darwin CJ, Carlyon RP (1995) in *Hearing: The Handbook of Perception and Cognition*, ed Moore BCJ (Academic, London), Vol 6, pp 387–424.
7. Shinn-Cunningham BG (2008) Object-based auditory and visual attention. *Trends Cogn Sci* 12:182–186.
8. Cusack R, Deeks J, Aikman G, Carlyon RP (2004) Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *J Exp Psychol Hum Percept Perform* 30:643–656.
9. Duncan J (1984) Selective attention and the organization of visual information. *J Exp Psychol Gen* 113:501–517.
10. Roelfsema PR, Lamme VAF, Spekreijse H (1998) Object-based attention in the primary visual cortex of the macaque monkey. *Nature* 395:376–381.
11. Serences JT, Liu T, Yantis S (2005) in *Neurobiology of Attention*, eds Itti L, Rees G, Tsotsos J (Academic, New York), pp 35–41.
12. VanRullen R, Carlson T, Cavanagh P (2007) The blinking spotlight of attention. *Proc Natl Acad Sci USA* 104:19204–19209.
13. Freyman RL, Helfer KS, McCall DD, Clifton RK (1999) The role of perceived spatial separation in the unmasking of speech. *J Acoust Soc Am* 106:3578–3588.
14. Shinn-Cunningham BG, Ihlefeld A, Satyavarta, Larson E (2005) Bottom-up and top-down influences on spatial unmasking. *Acust Acta Acust* 91:967–979.
15. Best V, Ozmeral E, Shinn-Cunningham BG (2007) Visually-guided attention enhances target identification in a complex auditory scene. *J Assoc Res Otolaryngol* 8:294–304.
16. Kidd G, Jr, Arbogast TL, Mason CR, Gallun FJ (2005) The advantage of knowing where to listen. *J Acoust Soc Am* 118:3804–3815.
17. Darwin CJ, Hukin RW (2000) Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *J Acoust Soc Am* 107:970–977.
18. Brungart DS, Simpson BD, Ericson MA, Scott KR (2001) Informational and energetic masking effects in the perception of multiple simultaneous talkers. *J Acoust Soc Am* 110:2527–2538.
19. Broadbent DE (1958) *Perception and Communication* (Pergamon, London).
20. Treisman AM (1971) Shifting attention between the ears. *Q J Exp Psychol* 23:157–167.
21. Brungart DS, Simpson BD (2007) Cocktail party listening in a dynamic multitalker environment. *Percept Psychophys* 69:79–91.
22. Arbogast TL, Kidd G, Jr (2000) Evidence for spatial tuning in informational masking using the probe-signal method. *J Acoust Soc Am* 108:1803–1810.
23. Mondor TA, Zatorre RJ (1995) Shifting and focusing auditory spatial attention. *J Exp Psychol Hum Percept Perform* 21:387–409.
24. Teder-Sälejärvi WA, Hillyard SA (1998) The gradient of spatial auditory attention in free field: An event-related potential study. *Percept Psychophys* 60:1228–1242.
25. Blaser E, Pylyshyn ZW, Holcombe A (2000) Tracking an object through feature-space. *Nature* 408:196–199.
26. Kidd G, Jr, Mason CR, Brughera A, Hartmann WM (2005) The role of reverberation in release from masking due to spatial separation of sources for speech identification. *Acust Acta Acust* 114:526–536.
27. Leonard RG (1984) A database for speaker-independent digit recognition. *Proceedings of the 1984 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '84)* (IEEE, Piscataway, NJ), Vol 9, pp 328–331.