مجله دانشکده پزشکی اصفهان
سال سی و یکم/شماره ۲۵۹/هفته اول دی ماه ۱۳۹۲

تاریخ دریافت: ۱۳۹۲/۶/۷
تاریخ پذیرش: ۱۳۹۲/۸/۲۴

بهبود وضوح گفتار در نویز با استفاده از الگوریتم ماسک باینری ایده‌آل

نادر ناصری[۱]، دکتر سعید کرمانی[۲]

مقاله پژوهشی

چکیده

مقدمه: کاربرد ماسک باینری ایده‌آل برای پردازش سیگنال گفتاری، بهبود قابل ملاحظه‌ای در وضوح گفتار هم در افراد با شنوایی طبیعی و هم افراد مبتلا به کم شنوایی نشان داده است. این ماسک به بخش‌های زمان- فرکانس سیگنال نویزی اعمال می‌گردد و بخش‌هایی از سیگنال پایین‌تر از سطح آستانه‌ی SNR (Signal-to-noise ratio) حذف می‌گردد و سایر بخش‌ها را عبور می‌دهد.

روش‌ها: در این مطالعه عوامل مؤثر بر روی الگوریتم ماسک باینری ایده‌آل مورد مطالعه و بررسی قرار گرفتند. تأثیر سطح آستانه SNR، سطح SNR ورودی، نوع ماسک کننده و تخمینگر نویز، بررسی و ارزیابی شد. تخمینگرهای جدیدی شامل وزنی Euclidean و COSH معرفی شدند. این تخمینگرها مبتنی بر درک سیستم شنوایی و ماسک شنیداری می‌باشند.

یافته‌ها: عملکرد بالای ماسک باینری در ناحیه‌ی ۱۵-۵ دسی‌بل مشاهده شد. یافته‌ها می‌تواند برای پیشرفت طراحی سمعک و پروتز کاشت حلزون مفید باشد.

نتیجه‌گیری: یافته‌های ما مؤید مطالعات پیشین در زمینه‌ی وضوح گفتار قابل توجه است؛ حتی زمانی که SNR، ۱۰- دسی‌بل باشد. ارزیابی عملکرد این الگوریتم نشان داد که تخمینگرهای جدید در مقایسه با تخمینگر Wiener می‌توانند حذف نویز بهتری داشته باشند.

واژگان کلیدی: به‌سازی گفتار، ماسک زمان- فرکانس، وضوح گفتار

ارجاع: ناصری نادر، کرمانی سعید. **بهبود وضوح گفتار در نویز با استفاده از الگوریتم ماسک باینری ایده‌آل.** مجله دانشکده پزشکی اصفهان ۱۳۹۲؛ ۳۱ (۲۵۹): ۱۷۹۶-۱۷۸۷

مقدمه

سیستم شنوایی انسان درک و جداسازی گفتاری کارآمدی در شرایط نویزی دارد. گر چه مکانیسم‌های اصلی به طور کامل درک نشده است اما آنالیز صحنه‌ی شنیداری (Auditory scene analysis یا ASA) به عنوان نظریه‌ی غالب در نظر گرفته می‌شود (۱). بر اساس این نظریه، شنودگان تفکیک گفتار را در دو مرحله انجام می‌دهند. در مرحله‌ی اول، ورودی اکوستیکی به قطعات زمان- فرکانس تجزیه می‌شود (۲، ۱). این قطعات در مرحله‌ی دوم با استفاده از الگوهای گروه‌بندی اولیه مثل متناوب بودن (Periodicity)، آغاز و پایان مشترک (Common onset/offset) دسته‌بندی می‌شوند. رویکردهای جداسازی گفتار و آنالیز

* این مقاله حاصل پایان‌نامه‌ی دوره‌ی کارشناسی ارشد به شماره‌ی ۳۹۲۲۷۶ در دانشگاه علوم پزشکی اصفهان است.

۱- دانشجوی کارشناسی ارشد، گروه فیزیک و مهندسی پزشکی، دانشکده‌ی پزشکی و کمیته‌ی تحقیقات دانشجویی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
۲ -استادیار، گروه فیزیک و مهندسی پزشکی، دانشکده‌ی پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران
نویسنده‌ی مسؤول: نادر ناصری
Email: nadernaseri.62@gmail.com

صحنه‌ی شنیداری محاسباتی (CASA یا Computational auditory scene analysis) مبتنی بر ASA هستند (۲). هدف اصلی CASA محاسبه‌ی ماسک باینری ایده‌آل (Ideal binary mask یا IBM) است که بر اساس پدیده‌ی ماسک شنیداری (Auditory masking) مطرح شده است (۳). IBM یک ماتریس باینری است که در حوزه‌ی زمان- فرکانس تعریف می‌شود. مقدار یک (متناظر با بخش زمان- فرکانس ماسک‌نشده) به معنای غالب بودن بخش زمان- فرکانس متناظر است، در حالی که مقدار صفر (واحد زمان- فرکانس ماسک‌شده) یعنی ماسک‌کننده (سیگنال نویزی) غالب است.

مطالعات متعددی به منظور بررسی تأثیر عوامل مختلف بر وضوح سیگنال‌های ماسک‌شده انجام شده‌اند (۸-۴). نتایج نشان می‌دهند که رویکرد IBM وضوح گفتار را در افراد با شنوایی طبیعی (۷، ۵-۴) و افراد مبتلا به نقص شنوایی (۷، ۶) به میزان چشمگیری بهبود می‌دهد. مشاهدات جالب دیگری در این آزمایشات صورت گرفت. Brungrat و همکاران از IBM جهت مطالعه‌ی تأثیر ماسک باینری در حضور گوینده‌های رقابتی (Competing talkers) استفاده کردند و مشاهده کردند که ناحیه‌ی با وضع ثابت (Plateau region) با وضوح بالا در معیار محلی (Local criterion) بین صفر تا ۱۲ دسی‌بل، قرار دارد (۴). نتایج Brungrat و همکاران (۴) و Wang و همکاران (۷) نشان دادند که معیار محلی ۶- دسی‌بل انتخاب بهتری است در صورتی که هدف بهبود وضوح سیگنال نویزی پردازش شده باشد. Li و Loizou ناحیه‌ی ثابت عملکردی وسیع‌تری در محدوده‌ی ۲۰-۵ دسی‌بل با استفاده از تبدیل فوریه گسسته (Discrete fourier transform) گزارش کردند و مشاهده کردند که این ناحیه‌ی خطی و رزولوشن بهتری در فرکانس‌های بالا دارد (۸). همچنین الگوریتم ماسک ایده‌آل در باز شناسی گفتار مورد بررسی قرار گرفته است (۱۱-۹).

به تازگی Roman و Woodurff نشان دادند که IBM می‌تواند وضوح گفتار را در شرایط نویزی و بازآوایی (Reverberant condition) بهبود ببخشد (۱۲).

در مطالعات پیشین، فرض می‌شد که ماسک باینری ایده‌آل موجود است. در عمل ماسک باینری باید از رویدادهای نویزی تخمین زده شود. این عمل در شرایط نویزی شدید دشوار است. در این پژوهش از تخمین‌گر بیزی استفاده شد که ماسک شنیداری ادراکی انسان به آن اضافه شده است، تا به تخمین و حذف نویز مؤثر دست یابیم. سایر عواملی که می‌تواند ماسک باینری را متأثر کنند شامل: انتخاب سطح آستانه، نوع ماسک‌کننده، مواد گفتاری و سطح SNR ورودی است. در این مطالعه وضوح ماسک باینری ایده‌آل با استفاده جملات IEEE (The Institute of Electrical and Electronics Engineers) به عنوان مواد آزمون و ماسک‌کننده (اصوات رقابتی) استفاده شد. در ادامه پیاده‌سازی ماسک باینری ایده‌آل و ارزیابی واقعی با استفاده مواد گفتاری بیان شد.

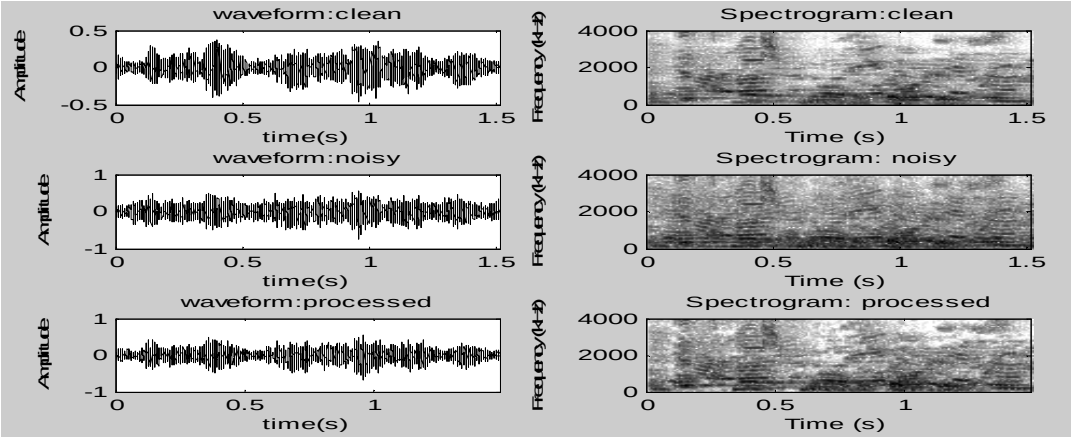
روش‌ها

مواد گفتاری مورد استفاده شامل مواد گفتاری برگرفته از پایگاه داده‌ی IEEE بود (۱۳). تمامی جملات به وسیله‌ی گویندگان مذکر تولید شدند. جملات مذکور در اتاقک اکوستیک با نرخ نمونه‌برداری ۲۵ کیلوهرتز

ضبط شدند. جزییات تجهیزات ضبط موجود است (۱۴). جملات با نویز همهمه (Babble) در مورد ۲۰ گوینده‌ی هم‌زمان ترکیب گردیدند. نویز همهمه در ۱۰۰ میلی‌ثانیه قبل از هر جمله شروع شد و ۱۰۰ میلی‌ثانیه پس از پایان هر جمله به اتمام رسید. تأثیر سطح آستانه‌ی SNR، سطح SNR ورودی، نوع ماسکر و تخمینگر نویز، بررسی و ارزیابی شد. تأثیر استفاده از انواع مختلف ماسک‌کننده نیز بررسی گردید.

پردازش سیگنال

پردازش شامل ساخت IBM بود که از سه نوع سیگنال هدف یا تمیز، ماسک‌کننده (نویز) و سیگنال مخلوط (Mixture) یا نویزی، استفاده شد. هر یک از این سیگنال‌ها در ابتدا با قطعات ۲۰ میلی‌ثانیه (با پنجره‌ی Hamming) و هم‌پوشانی ۵۰ درصد پنجره گذاری شدند. سپس تبدیل فوریه سریع (FFT یا Fast fourier transform) به هر قطعه اعمال شد. پس از تجزیه‌ی زمان- فرکانس، انرژی سیگنال هدف و ماسک‌کننده مقایسه شد. SNR محلی، در هر بخش زمان- فرکانس با مقدار سطح آستانه (Threshold) مقایسه گردید، تا مشخص شود که آن بخش زمان- فرکانس باید حفظ گردد (مقدار ماسک باینری یک است) یا حذف گردد (مقدار ماسک باینری صفر است). الگوی محاسبه‌شده مقادیر ماسک باینری شامل صفرها و یک‌ها به اندازه‌ی FFT طیف اصلاح‌شده‌ی IBM اعمال گردید. فاز طیف FFT مخلوط در معکوس FFT مورد استفاده قرار گرفت. محرک در نهایت در هر قطعه‌ی ۲۰ میلی‌ثانیه‌ای با به کارگیری روش Overlap and add ساخته شد. در این مطالعه، سطح آستانه از ۲۰-۱۰ دسی بل در گام‌های ۵ دسی‌بل تغییر کرد و عملکرد در هر مقدار آستانه مورد ارزیابی قرار گرفت. زمانی که آستانه‌ی SNR برابر صفر دسی‌بل بود، تنها واحدهای زمان- فرکانس سیگنال هدف که انرژی بالاتر از ماسک‌کننده داشتند حفظ شدند و مابقی بخش‌ها حذف گردیدند. شکل ۱ پردازش IBM برای مقدار آستانه‌ی ۱۰- دسی‌بل و سیگنال نویزی در SNR ۵- دسی‌بل را نشان می‌دهد.



شکل ۱. ردیف بالا طیف یک جمله از مواد گفتاری IEEE (The Institute of Electrical and Electronics Engineers) می‌باشد و همچنین همان سیگنال را در حوزه‌ی زمان نشان می‌دهد. ردیف وسط نشان‌دهنده‌ی طیف سیگنال نویزی است که با نویز Babble در SNR (Signal-to-noise ratio) ۵- دسی‌بل مخلوط شده است. ردیف پایین بیانگر سیگنال پردازش‌شده در سطح آستانه‌ی ۱۰- دسی‌بل می‌باشد.
پیکسل سفید مقدار یک و پیکسل سیاه مقدار صفر است.

سؤال مهم در این مطالعه تخمین نویز و نرخ توان سیگنال به نویز SNR بود و این که چه محدوده‌ای از آستانه‌های SNR بالاترین سطوح درک گفتار را فراهم می‌آورند.

تخمینگرهای بیزی عمومی

اگر $yn = x(n) + d(n)$ باشد که $y(n)$ سیگنال گفتار نویزی، شامل گفتار تمیز $x(n)$ و نویز $d(n)$ باشد. با گرفتن تبدیل فوریه از $y(n)$ داریم (معادله‌ی ۱):

معادله‌ی ۱. $Y(\omega_K) = X(\omega_K) + D(\omega_K)$

در این مطالعه، علاقه‌مند بودیم که اندازه‌ی طیف X_k را از طیف $Y(\omega_K)$ تخمین بزنیم. فرض می‌کنیم $\varepsilon = X_k - \hat{X}_k$ خطای تخمین اندازه در فرکانس kام باشد و فرض می‌کنیم که $d(\varepsilon) \hat{=} d\left(X_k, \hat{X}_k\right)$ تابع غیر منفی بر اساس ε باشد امید ریاضی تابع هزینه $E\left[d\left(X_k, \hat{X}_k\right)\right]$ بر حسب pdf توأم $(p(X_k, y(\omega_K))$ تابع هزینه‌ی بیز $\Re$ نامیده می‌شود و به صورت زیر است (معادله‌ی ۲):

معادله‌ی ۲.

$$\begin{aligned}\Re &= E[d(X_k, \hat{X}_k)] \\ &= \iint d(X_k, \hat{X}_k)\, p(X_k, Y(\omega_k))dX_k dY(\omega_k) \\ &= \int \left[\int d(X_k, \hat{X}_k) p(X_k \mid Y(\omega_k))dX_k\right] p(Y(\omega_k))dY(\omega_k)\end{aligned}$$

مینیمم کردن تابع بیز $\Re$ بر حسب $\hat{X}_k$ برای تابع هزینه‌ی دلخواه، تخمینگرهای مختلفی را نتیجه می‌دهد (۱۵).

به طور خلاصه، تخمینگرهای بیزی مختلف X_k می‌تواند بر اساس انتخاب تابع هزینه به دست آیند. با توجه به توابع هزینه‌ی مورد استفاده در مطالعات Wolfe و Godsill (۱۶)، Lotter و Vary (۱۷) و Plourde و Champagne (۱۸) و تابع هزینه‌ی خطای مربعات توابع، لازم نیست نوع تابع هزینه به صورت ذهنی معنی‌دار باشد (۱۸). در ادامه تخمینگرهای بیزی X_k مبتنی بر توابع هزینه‌ی ادراکی به جای تابع هزینه‌ی مربع خطا استفاده می‌شود. این توابع هزینه، جهت تخمین نویز استفاده می‌شوند.

تخمینگر Euclidean وزنی

برخی از معیارهای فاصله‌ای مثل معیار Itakura-Saito بیش از سایر معیارها مبتنی بر ادراک هستند. می‌توان از دیگر معیار فاصله‌ای معنی‌دار مثل تخمین Euclidean وزنی (Weighted euclidean estimator) استفاده کرد. تابع هزینه بر اساس معیار خطای ورنی ادراکی می‌باشد. تابع هزینه‌ی تخمینگر Euclidean وزنی به شکل زیر است (معادله‌ی ۳).

معادله‌ی ۳. $d_{we}(X_k, \hat{X}_k) = X_k^p (X_k, \hat{X}_k)^2$

P یک مقدار حقیقی است. زمانی که از معادله‌ی ۳ را در معادله‌ی ۲ استفاده کنیم معادله‌ی ۴ به دست می‌آید.

معادله‌ی ۴. $\hat{X}_k = \frac{\int_0^\infty x_k^{p+1} p(X_k|Y(\omega_k))dX_k}{\int_0^\infty x_k^{p} p(X_k|Y(\omega_k))dX_k}$

بهره‌ی تخمینگر Euclidean وزنی به صورت معادله‌ی ۵ به دست می‌آید.

معادله‌ی ۵.

$$\hat{X}_k = \frac{\sqrt{V_k}}{\gamma_k} Y_k \frac{\Gamma\left(\frac{p+1}{2}+1\right)\Phi\left(\frac{p+1}{2}, 1; -V_k\right)}{\Gamma\left(\frac{p}{2}+1\right)\Phi\left(-\frac{p}{2}, 1; -V_k\right)} Y$$

$p > -2$

که برای مقادیر $p > -2$ $p = -1$ در نظر گرفته می‌شود.

این تخمینگر از مزیت ویژگی ماسک شنیداری گوش انسان برخوردار است. مقدار p تعادل بین اعوجاج گفتاری و کاهش نویز برقرار می‌کند. مقدار $p = -1$ این تعادل را به خوبی برقرار می‌کند.

تخمینگر COSH

تخمینگر COSH مبتنی بر معیار SaitoItakura-

می‌باشد و حالت متقارن آن است. تابع هزینه‌ی مورد استفاده‌ی COSH در تخمین بیز به صورت معادله‌ی ۶ در نظر گرفته می‌شود.

معادله‌ی ۶.

$$d_{\cosh}(X_K,\hat{X}_k)=\cosh\left(\log\frac{X_k}{\hat{X}_k}\right)-1=\frac{1}{2}\left[\frac{X_k}{\hat{X}_k}+\frac{\hat{X}_k}{X_k}\right]-1$$

هنگامی که از معادله‌ی ۶ در معادله‌ی ۲ استفاده کنیم معادله‌ی ۷ به دست می‌آید.

معادله ی ۷.

$$\hat{X}_k^2=\frac{\int_0^\infty x_k^{p+1}p(X_k|Y(\omega_k))dX_k}{\int_0^\infty x_k^{p-1}p(X_k|Y(\omega_k))dX_k}$$

بهره‌ی متناظر تخمینگر COSH با معادله‌ی ۸ محاسبه می‌گردد.

معادله‌ی ۸.

$$\hat{X}_k=\frac{1}{\gamma_k}\sqrt{\frac{v_k\Gamma\left(\frac{p+3}{2}\right)\Phi\left(-\frac{p+1}{2},1;-v_k\right)}{\Gamma\left(\frac{p}{2}+1\right)\Phi\left(-\frac{p-1}{2},1;-v_k\right)}}Y$$

به طور مشابه مشاهده می‌شود که مقدار p توازن بین اعوجاج گفتاری و حذف نویز را برقرار می‌کند.

ارزیابی عملکرد

برای ارزیابی الگوریتم از دو روش ارزیابی واقعی، SNR قطعه‌ای و معیار ارزیابی ادراکی کیفیت گفتار (Perceptual evaluation of speech quality یا PESQ) استفاده کردیم. SNR قطعه‌ای، مبتنی بر SNR کلاسیک است. یکی از روش‌های مرسوم برای ارزیابی الگوریتم‌ها، به‌سازی گفتار است. چون همبستگی SNR کلاسیک و کیفیت ذهنی پایین است. SNR قطعه‌ای را انتخاب کردیم که با میانگین‌گیری فریم‌های گفتاری سطح SNR به دست می‌آید (معادله‌ی ۹).

معادله‌ی ۹.

$$\text{segSNR}_{dB}=\frac{1}{M}\sum_{m=1}^{M-1}10\log\left[\frac{\sum_{n=0}^{L-1}|s(n+mL)|^2}{\sum_{n=0}^{L-1}|s(n+mL)-\hat{s}(n=mL)|^2}\right]$$

s(n) سیگنال گفتاری تمیز و $\hat{S}(n)$ سیگنال بهبودیافته است. M تعداد فریم‌ها و L تعداد نمونه‌ها در هر فریم است. سطوح بالا و پایین آستانه‌ها به ترتیب +۳۵ دسی‌بل و ۱۰- دسی‌بل است.

Perceptual evaluation of speech quality (PESQ) یکی از معمول‌ترین روش‌های پیش‌بینی کیفیت گفتار است (۱۹). در معیار PESQ، سیگنال رفرنس و سیگنال بهبودیافته (پردازش‌شده) در ابتدا از لحاظ زمانی و سطحی هم‌تراز شدند. سپس با استفاده از تبدیلات درکی مهم پردازش شدند. پردازش شامل آنالیز طیفی بارک، همسان‌سازی فرکانس، همسان‌سازی بهره و بلندی بود. اختلاف یا در اصطلاح تفاضل بین طیف بلندی در حوزه‌ی زمان و فرکانس برای پیش‌بینی کیفیت گفتار انجام می‌شود. امتیاز PESQ از مقدار یک (بدترین) تا مقدار ۵/۴ (بهترین) می‌باشد و امتیازات بالاتر بیانگر کیفیت بهتر هستند.

یافته‌ها

بررسی تأثیرات سطح آستانه SNR و سطح SNR ورودی

سطح آستانه (T) از ۲۰- تا +۱۰ دسی‌بل در گام‌های ۵ دسی‌بل تغییر می‌کند. نتایج PESQ و SNR قطعه‌ای در شکل ۲ نشان داده شده‌اند. پردازش شامل ۱۴ حالت (۱ ماسک‌کننده × ۷ سطح آستانه × ۲ سطح SNR) بود. محدوده‌ی مقادیر آستانه‌ای برای عملکرد ثابت در SNR ۱۰- دسی‌بل پایین‌تر از SNR ۵- دسی‌بل بود. این دو ارزیابی واقعی نشان می‌دهند که ناحیه‌ی با عملکرد ثابت در مورد SNR ۱۰- دسی‌بل، ناحیه‌ی ۱۵- دسی‌بل

تا صفر دسی‌بل بود. در حالی که در مورد SNR ۵- دسی‌بل، ناحیه‌ی ۱۵- تا ۵ دسی‌بل بود. واحدهای زمان– فرکانس در سطوح پایین‌تر از ۱۵- دسی‌بل بین سیگنال هدف و سیگنال نویز قابل تشخیص نیستند.

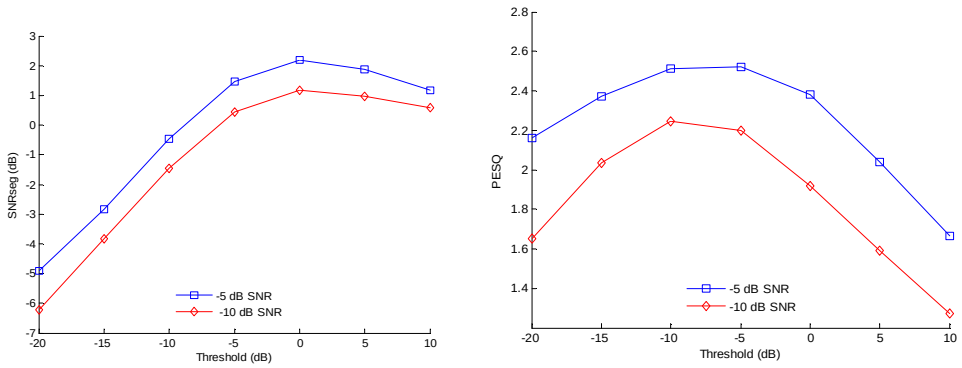
بررسی تأثیرات نوع ماسک‌کننده

در بررسی پیشین، عملکرد را با استفاده از یک نوع ماسک‌کننده (نویز همهمه) بررسی کردیم. این عملکرد می‌تواند با انواع مختلف ماسک‌کننده متأثر گردد. در این قسمت عملکرد الگوریتم IBM را با استفاده از نویز حالت ثابت (Steady state noise) و نویز همهمه (Babble) ودو گوینده‌ی رقابتی (Two talker) بررسی کردیم. این بررسی نشان داد که الگوریتم IBM زمانی که ماسک اجزای داده‌ای (زمانی که ماسک‌کننده گوینده‌ی رقابتی بود) را دارد نسبت به ماسک انرژی (زمانی که ماسک‌کننده به طور کامل مبتنی بر انرژی بود مانند مانند نویز حالت ثابت) مؤثرتر بود. در این مطالعه سه نوع ماسک‌کننده استفاده شدند. اولی نویز حالت ثابت (Simultaneous switching noise یا SSN) بود که طیف ثابتی در جملات مورد آزمایش مواد گفتاری IEEE داشت. دومی نویز همهمه بود (گویندگان رقابتی) و سومی دو گوینده‌ی هم‌زمان بودند. پردازش شامل ۲۱ حالت (۳ ماسک‌کننده × ۷ سطح آستانه) است. سطح آستانه‌ی مورد آزمایش از ۲۰- تا ۱۰ دسی‌بل در گام‌های ۵ دسی‌بل می‌باشد. ۳۰ جمله از جملات مواد گفتاری IEEE در هر حالت پردازشی مورد استفاده قرار می‌گیرند. میانگین معیار SNR قطعه‌ای و PESQ در شکل ۳ نشان داده شده‌اند. محدوده‌ی عملکردی با وضع ثابت در هر سه نوع ماسک‌کننده از ۱۵- تا +۵ دسی‌بل بود. به نظر می‌رسد که الگوریتم IBM در مورد وضوح گفتار، زمانی که گفتار به وسیله‌ی نویز گفتاری (گوینده‌ی رقابتی) ماسک می‌شود از گفتار ماسک‌شده با نویز، عملکرد بهتری داشت.

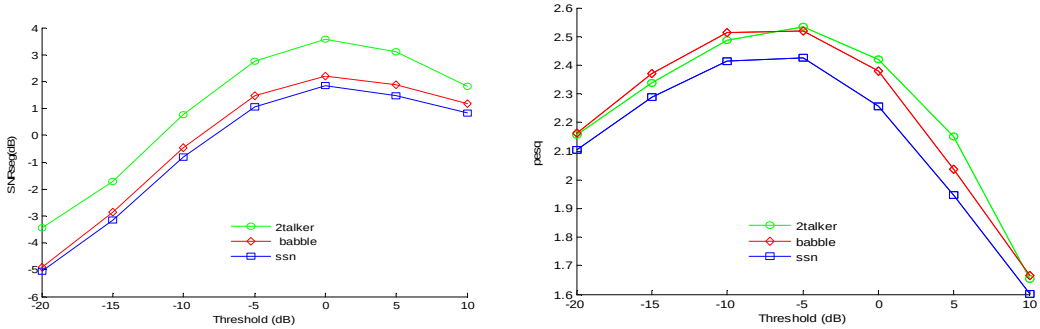
بررسی تأثیر تخمینگرهای نویز

این بخش ارزیابی و مقایسه‌ی عملکرد تخمینگرهای نویز را ارائه می‌دهد. اولین تخمینگر Euclidean وزنی، دومین تخمینگر COSH و سومین تخمینگر wiener بودند. مواد گفتاری استفاده‌شده در این بخش جملات برگرفته از پایگاه داده‌ی Noizeus (۱۴)، شامل سیگنال‌های فیلترشده با پهنای باند فرکانسی تلفنی بود که با نرخ ۸ کیلوهرتز نمونه‌برداری شد. نویزهای مورد استفاده، نویز خیابان (Street noise) و نویز اتومبیل (Car noise) برگرفته از پایگاه داده‌ی Noizeus بودند. پردازش شامل ۲۴ حالت (۲ ماسک‌کننده × ۴ سطح SNR × ۴ الگوریتم مختلف) بود. برای ارزیابی واقعی الگوریتم همانند قبل معیار PESQ و معیار SNR قطعه‌ای استفاده شد. همان‌طور که مشاهده می‌شود تخمینگرهای پیشنهادی بر حسب معیارهای SNR قطعه‌ای PESQ و سطوح آستانه از ۱۰- تا +۱۰ دسی‌بل در گام‌های ۵ دسی‌بل ارزیابی شدند.

با توجه به نتایج واقعی در جدول ۱ مشاهده می‌شود که تخمینگر Euclidean وزن داده‌شده، امتیازات عملکردی بالاتری نسبت به سایر تخمینگرها برای سطوح مختلف SNR داشت. تخمینگر Euclidean وزن داده‌شده اعوجاج نویزی پایین‌تری نسبت به تخمینگر Wiener و تخمینگر COSH دارد.



شکل ۲. عملکرد بر حسب میانگین PESQ (Perceptual evaluation of speech quality) و میانگین SNR (Signal-to-noise ratio) قطعه‌ای به صورت تابعی از سطح آستانه‌ی SNR برای دو مقدار SNR، ماسک‌کننده از نویز همهمه (Babble) می باشد.



شکل ۳. عملکرد بر حسب میانگین معیار PESQ (Perceptual evaluation of speech quality) و میانگین SNR (Signal-to-noise ratio) قطعه‌ای به صورت تابعی از سطح آستانه‌ی SNR برای سه نوع ماسک‌کننده: نویز همهمه (Babble)، نویز حالت ثابت (SSN یا Simultaneous switching noise) و دو گوینده‌ی رقابتی (Two talker). ارزیابی با استفاده از سیگنال نویزی

جدول ۱. امتیازات کیفیت واقعی برای الگوریتم‌های مختلف: برای نویز اتومبیل (Car noise) و نویز خیابان (Street noise) در SNR صفر دسی‌بل و SNR ۵ دسی‌بل)

Noise	روش	SNR ۵ دسی‌بل		SNR صفر دسی‌بل	
		PESQ	SNRseg	PESQ	SNRseg
اتومبیل	Welucid	۲/۳۲۹۳	۱/۶۸۴۸	۱/۹۵۶۸	-۰/۶۴۱۴
	Wcosh	۲/۳۰۳۷	۱/۵۸۶۷	۱/۹۳۸۴	-۰/۶۶۷۱
	Wiener	۲/۲۲۶۹	۰/۳۸۲۲	۱/۹۱۲۱	-۱/۸۳۵۰
	Noisy	۱/۸۹۱۳	۲/۲۵۳۱	۱/۶۳۳۷	-۴/۹۸۷۱
خیابان	Welucid	۲/۲۰۴۴	۰/۷۸۵۶	۱/۸۳۵۳	-۳۰۱۹/۱
	Wcosh	۲/۱۴۵۰	۰/۸۰۰۴	۱/۷۸۳۲	-۱/۲۷۷۱
	Wiener	۲/۱۴۵۴	۰/۰۸۹۲	۱/۸۰۰۷	-۱/۱۴۵۴
	Noisy	۱/۹۰۴۴	-۱/۶۷۸۰	۱/۵۶۳۰	-۴/۳۰۱۷

SNR: Signal-to-noise ratio; PESQ: Perceptual evaluation of speech quality

بحث

نتایج بررسی سطح آستانه و نوع ماسک‌کننده نشان‌دهنده‌ی این بود که درک گفتار در محیط‌هایی که چند گوینده‌ی هم‌زمان دارند بهبود قابل ملاحظه‌ای می‌یابند. مطالعه‌ی حاضر تاییدگر یافته‌های سایر مطالعات می‌باشد (۵-۴). ناحیه‌ی به دست‌آمده با وضع ثابت (نزدیک ۱۰۰ درصد) در ناحیه‌ی نزدیک به SNR صفر دسی‌بل بود و برای مقادیر بالا و پایین آستانه SNR محلی کاهش یافت. این الگو برای تمامی ماسک‌کننده‌های مورد آزمایش چه نویز گفتاری و یا نویز حالت ثابت، یکسان بود.

الگوریتم IBM زمانی که گفتار به وسیله‌ی نویز گفتاری (که به طور عمده ماسک داده‌ای است) نسبت به نویز حالت ثابت (به طور کامل ماسک انرژی است) باشد، از لحاظ بهبود کیفیت گفتار کارآمدتر است. بررسی تأثیر سطح آستانه نشان کرد که افزایش ۱ دسی‌بل در آستانه‌ی SNR، به همان اندازه واحدهای زمان- فرکانس را حذف می‌کند. اگر SNR کلی ۱ دسی‌بل فزایش داشته باشد. ناحیه‌ی با وضع ثابت در مطالعه‌ی Brungart و همکاران ۱۲-۰ دسی‌بل بود (۴)، در حالی که در مطالعه‌ی حاضر در اغلب موارد ۱۵-۰ دسی‌بل بود که نزدیک به ناحیه‌ی با وضع ثابت مطالعه‌ی Brungart و همکاران بود. به علاوه مواد آزمون گفتاری آن‌ها، CRM (Coordinated response measure) بود که از لحاظ محتوا متفاوت است.

این مقاله الگوریتم ماسک باینری ایده‌آل را پیاده‌سازی و ارزیابی کرده است در حالت کلی نتایج بررسی‌ها مؤید مطالعات پیشین بود. که ماسک اطلاعات را در درک گفتار چند گوینده بررسی کرده بودند و نشان داد که ماسک داده‌ای بر درک چند گوینده غالب است و ماسک انرژی تأثیر بالاتری بر روی گفتار در نویز نسبت به ماسک گفتار بر گفتار دارد. این نتایج کاربرد قوی CASA را نشان می‌دهد. IBM به عنوان هدف محاسباتی CASA معرفی شد و نتایج بررسی‌ها نشان داد که IBM تکنیک مؤثری در بهبود گفتار است، زمانی که گویندگان رقابتی وجود داشته باشند.. در بررسی تخمینگرهای نویز با استفاده از معیارهای واقعی مشاهده شد که تخمینگر Euclidean وزن داده‌شده عملکرد بهتری داشت و تعادل بهتری بین مقدار کاهش نویز و اعوجاج گفتار و سطح نویز باقی‌مانده موزیکال نسبت به سایر تخمینگرها داشت.

الگوریتم IBM می‌تواند جهت به‌سازی گفتار، حذف نویز در سمعک و پروتز کاشت حلزون استفاده گردد.

References

1. Bregman AS. Auditory Scene Analysis: Hearing in Complex Environments. In: McAdams S, Bigand E, Editors. Thinking in Sound: The Cognitive Psychology of Human Audition. Oxford, UK: Oxford University Press; 1993. p. 10-36.
2. Wang D, Brown GJ. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. New Jersey, NJ: Wiley; 2006.
3. Wang D. On Ideal Binary Mask As the Computational Goal of Auditory Scene Analysis. Speech Separation by Humans and Machines 2005; 181-97.
4. Brungart DS, Chang PS, Simpson BD, Wang D. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. J Acoust Soc Am 2006; 120(6): 4007-18.
5. Cao S, Li L, Wu X. Improvement of intelligibility of ideal binary-masked noisy

speech by adding background noise. J Acoust Soc Am 2011; 129(4): 2227-36.
6. Anzalone MC, Calandruccio L, Doherty KA, Carney LH. Determination of the potential benefit of time-frequency gain manipulation. Ear Hear 2006; 27(5): 480-92.
7. Wang D, Kjems U, Pedersen MS, Boldt JB, Lunner T. Speech intelligibility in background noise with ideal binary time-frequency masking. J Acoust Soc Am 2009; 125(4): 2336-47.
8. Li N, Loizou PC. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. J Acoust Soc Am 2008; 123(3): 1673-82.
9. Hartmann W, Fosler-Lussier E. Investigations into the incorporation of the Ideal Binary Mask in ASR. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2011 May 22-27 May; Prague, Czech Republic; 2011.
10. De Souza Siqueira Versiani T, Rodrigues GF, de Souza ACS, de Matos Moreira J, Yehia HC. Binary spectral masking for speech recognition systems. Proceedings of the 35th International Conference on Telecommunications and Signal Processing (TSP); 2012 Jul 3-4; Prague, Czech Republic; 2012.
11. Ahmadi M, Gross VL, Sinex DG. Perceptual learning for speech in noise after application of binary time-frequency masks. J Acoust Soc Am 2013; 133(3): 1687-92.
12. Roman N, Woodruff J. Intelligibility of reverberant noisy speech with ideal binary masking. J Acoust Soc Am 2011; 130(4): 2153-61.
13. Rothauser EH, Chapman WD, Guttman N, Hecker MH., Nordby KS, Silbiger HR, et al. IEEE recommended practice for speech quality measurements. IEEE Transactions on Audio and Electroacoustics 1969; 17(3): 225-46.
14. Hu Y, Loizou PC. Subjective comparison and evaluation of speech enhancement algorithms. Speech communication 2007; 49(7): 588-601.
15. Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. Acoustics, Speech and Signal Processing, IEEE Transactions on 1984; 32(6): 1109-21.
16. Wolfe PJ, Godsill SJ. Towards a perceptually optimal spectral amplitude estimator for audio signal enhancement. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing; 2000 Jun 5-9; Istanbul, Turkey; 2000.
17. Lotter T, Vary P. Speech enhancement by map spectral amplitude estimation using a super-Gaussian speech model. EURASIP Journal on Applied Signal Processing 2005; 2005: 1110-26.
18. Plourde E, Champagne B. Auditory-Based Spectral Amplitude Estimators for Speech Enhancement. Audio, Speech, and Language Processing, IEEE Transactions on 2008; 16(8): 1614-23.
19. Rix AW, Beerends JG, Hollier MP, Hekstra AP. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing; 2001 May 7-11; Salt Lake City, UT; 2001.

Journal of Isfahan Medical School
Received: 29.08.2013
Vol. 31, No. 259, 1st Week, January 2014
Accepted: 15.11.2013

Improving Speech Intelligibility Using Ideal Binary Mask

Nader Naseri[1], Saeid Kermani PhD[2]

Original Article

Abstract

Background: The application of the ideal binary mask (IBM) for speech signal processing provides remarkable intelligibility improvements in both normal-hearing and hearing-impaired listeners. Binary mask widely applies to the time-frequency (T–F) representation of a noisy signal and eliminates units of a signal below a signal-to-noise-ratio (SNR) threshold while retains others.

Methods: The factors underlying intelligibility of ideal binary-masked speech were examined and evaluated in the present study. The effects of the local SNR threshold, input SNR level, masker type, and ideal mask-estimator were examined. New estimators including weighted Euclidean and COSH were proposed in which, the human perceptual auditory masking effect and perceptual perception were incorporated.

Findings: High-performance plateau for SNR thresholds ranging from −20 to 5 dB was observed. Findings could be used for hearing-aid and cochlear-implant designs.

Conclusion: Intelligibility of speech was high even at −10 dB SNR for all maskers tested. Performance assessment shows that our proposed estimators can achieve more significant noise estimation as compared to the Wiener estimator.

Keywords: Speech enhancement, Binary masking, Speech intelligibility

Citation: Naseri N, Kermani S. **Improving Speech Intelligibility Using Ideal Binary Mask.** J Isfahan Med Sch 2014; 31(259): 1787-96

* This paper is derived from a MSc thesis No. 392276 in Isfahan University of Medical Sciences.

1- MSc Student, Department of Medical Physics and Medical Engineering, School of Medicine AND Student Research Committee, Isfahan University of Medical Sciences, Isfahan, Iran
2- Assistant Professor, Department of Medical Physics and Medical Engineering, School of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran
Corresponding Author: Nader Naseri, Email: nadernaseri.62@gmail.com