تخمین SNR ورودی با استفاده از ماسک باینری در سیستم‌های مبتنی بر آنالیز ترکیب شنیداری محاسباتی

مسعود گراوانچی‌زاده[۱]، دانشیار؛ پریا دادور[۲]، کارشناس‌ارشد

۱- دانشکده مهندسی برق و کامپیوتر - دانشگاه تبریز - تبریز - ایران - geravanchizadeh@tabrizu.ac.ir
۲- دانشکده مهندسی برق و کامپیوتر - دانشگاه تبریز - تبریز - ایران - paria.dadvar88@ms.tabrizu.ac.ir

چکیده: در این مقاله، روش جدیدی برای تخمین نسبت سیگنال به نویز (SNR) سیگنال ترکیب ارائه شده است که بر پایه روش آنالیز ترکیب شنیداری محاسباتی (CASA) است. در روش ارائه‌شده، ماسک باینری ایده‌آل (IBM) که به‌طور معمول هدف محاسباتی سیستم‌های مبتنی بر CASA است، برای تخمین SNR سیگنال گفتار نویزی به کار گرفته می‌شود. روش پیشنهادی با استفاده از IBM و چندین ماسک شبه IBM ارزیابی شده است. این روش، ساده و از نظر محاسباتی کارآمد است. ارزیابی‌های اصولی نشان می‌دهند که روش پیشنهادی، در محدوده وسیعی از مقادیر SNR، تخمین قابل‌قبولی از سطح SNR ورودی ارائه می‌دهد و خطاهای احتمالی در تخمین IBM عملکرد سیستم پیشنهادی را چندان تحت‌تأثیر قرار نمی‌دهد. نتایج شبیه‌سازی‌ها همچنین نشان می‌دهند که سیستم پیشنهادی عملکردی بهتری را نسبت به روش‌های قبلی تخمین SNR دارد.

واژه‌های کلیدی: تخمین SNR، نسبت سیگنال به نویز، آنالیز ترکیب شنیداری محاسباتی (CASA)، ماسک باینری ایده‌آل (IBM).

Input SNR Estimation using Binary Mask in Systems based on Computational Auditory Scene Analysis

M. Geravanchizadeh[1], Associate Professor; P. Dadvar[2], MSc

1- Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, Email: geravanchizadeh@tabrizu.ac.ir
2- Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, Email: paria.dadvar88@ms.tabrizu.ac.ir

Abstract: This paper presents a new approach for estimating the signal-to-noise ratio (SNR) of mixture signal, which is based on the computational auditory scene analysis (CASA). The ideal binary mask (IBM) which is generally the computational goal of CASA-based systems is used to estimate the SNR of noisy speech signal. The proposed method is evaluated using IBM and some quasi-IBM masks. The method is simple and computationally efficient. Systematic evaluations show that the proposed method results in reasonable estimation of the input SNR level in a wide range of SNR values, and probable errors in estimating IBM do not affect so much the performance of the proposed system. Furthermore, simulation results show that the proposed system outperforms previous SNR estimation methods.

Key Words: SNR estimation, signal-to-noise ratio, computational auditory scene analysis (CASA), Ideal binary mask (IBM).

تاریخ ارسال مقاله: ۱۳۹۳/۰۲/۱۶
تاریخ اصلاح مقاله: ۱۳۹۳/۰۸/۲۶
تاریخ پذیرش مقاله: ۱۳۹۴/۰۱/۲۹
نام نویسنده مسئول: مسعود گراوانچی‌زاده
نشانی نویسنده مسئول: ایران - تبریز - بلوار ۲۹ بهمن - دانشگاه تبریز - دانشکده مهندسی برق و کامپیوتر.

۱- مقدمه

بخش قابل‌توجهی از اطلاعاتی که توسط انسان‌ها از جهان پیرامون دریافت می‌شود، توسط حس شنیداری به دست می‌آید. کاربردها و محصولات تکنولوژی گفتاری مختلفی مانند سیستم‌های بهبود گفتار[۱] وجود دارند که با ورودی گفتار در شرایط نویزی کار می‌کنند. طراحی و توسعه سیستم‌های پایدار در برابر نویز سبب شده است تا نیاز به آنالیز عملکرد الگوریتم‌های موجود در حضور نویز پس‌زمینه[۲] با سطوح مختلف به وجود آید. در بسیاری از کاربردهای پردازش سیگنال گفتار دیجیتال، یک تخمین مناسب از نسبت سیگنال به نویز[۳] (SNR) محیط صوتی موردنیاز است. به‌عنوان‌مثال، در الگوریتم‌های فشرده‌سازی[۴] و تقویت‌کنندگی[۵] می‌توان با انتخاب بهینه پارامترها بر اساس SNR، عملکرد این سیستم‌ها را ارتقاء بخشید. در کاربرد دیگر، تخمین SNR می‌تواند به طور مستقیم به‌عنوان پارامتر کنترل برای الگوریتم‌های کاهش نویز (مانند فیلتر وینر[۶] یا روش کاهش طیفی[۷]) موردستفاده قرار گیرد.

تخمین SNR ورودی مسئله چالش‌انگیزی است. به‌عنوان‌مثال، بیش‌تر روش‌های تخمین SNR حاضر، از آشکارسازی فعالیت صدا[۸] (VAD) بهره می‌گیرند و نتیجه این روش‌ها به عملکرد VAD پیاده‌سازی‌شده بستگی شدید دارد [۱]. بااین‌وجود، روش‌هایی وجود دارند که نیازی به پیاده‌سازی VAD ندارند. به‌عنوان‌مثال، تخمین SNR بلندمدت[۹]، در باندهای فرکانسی تکی در [۲] ارائه شده است. در این مرجع، یک تخمین‌گر SNR ادراکی مبتنی بر ویژگی که از مدولاسیون‌های طیفی-زمانی استفاده می‌کند، ارائه شده است. بااین‌وجود، فاز آموزشی[۱۰] برای شبکه‌های عصبی پیاده‌سازی‌شده در سیستم، بار محاسباتی قابل‌توجهی دارد. در الگوریتم‌های تخمین SNR کوتاه‌مدت[۱۱]، تخمین سطح نویز مسئله مهمی است و به‌طور گسترده موردمطالعه قرار گرفته است.

ازجمله این پژوهش‌ها، روش مبتنی بر هیستوگرام طیفی Hirsch [۳] و روش ردیابی پوش کم‌انرژی[۱۲] Martin [۴] است. SNR پیشین[۱۳]، که نسبت توان سیگنال به نویز است، به‌طور گسترده در الگوریتم‌های بهبود گفتار به کار رفته است و معمولاً با استفاده از روش تصمیم جهت‌دار[۱۴] Ephraim و Malah [۵] تخمین زده می‌شود. تخمین SNR کل گفتار کم‌تر از تخمین SNR محلی[۱۵] موردمطالعه قرار گرفته است. الگوریتم SPQA[۱۶] از مؤسسه ملی استانداردها و فناوری (NIST)[۱۷] [۶]، با استفاده از گفتار نویزی، نمایش هیستوگرامی از توان سیگنال کوتاه‌مدت می‌سازد که برای به دست آوردن توزیع‌های گفتار نویزی و نویز مورداستفاده قرار می‌گیرد. با استفاده از این توزیع‌ها، نسبت سیگنال به نویز بیشینه[۱۸] به جای SNR میانگین محاسبه می‌شود. SNR بیشینه به وضوح، تخمین دست بالایی[۱۹] از SNR واقعی است. Dat و همکاران [۷] از روش مشابهی استفاده کرده‌اند، ولی آن‌ها به جای تطبیق یک هیستوگرام، با استفاده از الگوریتم بیشینه‌سازی انتظار[۲۰] (EM)، یک گوسی دو جزئی[۲۱] را با داده تطبیق داده‌اند. Dat و همکاران، با استفاده از گوسی‌های آموزش داده‌شده[۲۲] یک روش اصولی برای استخراج SNR سیگنال ارائه کرده‌اند. مشابه با مرجع [۱]، روش آن‌ها زمانی که فرض گوسی دووجهی[۲۳] برآورده نمی‌شود، دچار مشکل می‌شود.

روش Kim و Stern [۸] بر اساس توزیع دامنه شکل موج است. این روش فرض می‌کند که گفتار تمیز و نویزی، دارای توزیع گاما[۲۴] هستند و نویز دارای توزیع گوسی است. این روش، SNR کلی را بر اساس تخمین پارامترهای توزیع با استفاده از گفتار نویزی، تخمین می‌زند. الگوریتم آن‌ها زمانی که این فرضیات برآورده شوند، عملکرد خوبی دارد و اگرچه با فرض گوسی بودن نویز طراحی شده است، برای نویزهای موسیقی و گفتار نیز به خوبی کار می‌کند و نتایج بهتری را نسبت به مرجع [۱] نشان می‌دهد. کاهش عملکرد در شرایط SNR پایین و زمانی که نویز پس‌زمینه مشخصه‌های غیر گوسی دارد، رخ می‌دهد. یک روش جایگزین نسبتاً ساده، روش ردیابی چگالی طیفی توان نویز مبتنی بر کمینه خطای مجذور میانگین[۲۵]، ارائه‌شده توسط Hendriks و همکاران، است که از الگوریتم‌های بهبود گفتار برای تخمین چگالی طیفی توان نویز [۹] و مربع اندازه گفتار در حوزه DFT [۱۰] استفاده می‌کند. با فرض اینکه PSD نویز، انرژی نویز را تقریب می‌زند، که فرضی منطقی است، SNR کلی پهن‌باند[۲۶]، با جمع زدن این تخمین‌ها در زمان، مستقیماً قابل محاسبه است.

اخیراً، تلاش‌های فراوانی در زمینه آنالیز ترکیب شنیداری محاسباتی[۲۷] (CASA) به‌منظور جداسازی گفتار از ترکیب‌های نویزی انجام شده است [۱۱، ۱۲]. به‌علاوه، پژوهشگران بسیاری با الهام از پدیده پوشانندگی شنیداری[۲۸]، ماسک باینری ایده‌آل[۲۹] (IBM) را به‌عنوان هدف محاسباتی CASA پذیرفته‌اند [۱۳، ۱۴].

این مقاله روشی ساده و در عین حال کارآمد برای تخمین SNR ورودی در سیستم‌های مبتنی بر آنالیز ترکیب شنیداری محاسباتی ارائه می‌دهد. این تخمین با استفاده از یک ماسک باینری ایده‌آل (IBM) یا یک IBM تخمین‌زده‌شده[۳۰] انجام می‌شود و می‌تواند در سیستم‌های مبتنی بر CASA برای ارتقاء عملکرد این سیستم‌ها به کار رود. سیستم پیشنهادی، با فرضیات زیر SNR گفتار نویزی را با استفاده از یک ماسک باینری تخمین می‌زند: اول اینکه، انرژی کل گفتار تقریباً برابر با انرژی کل سیگنال نویزی در واحدهای زمان- فرکانس[۳۱] (T-F) پوشانده‌نشده غالباً گفتار[۳۲] (یک‌ها در IBM) بوده و دوم اینکه، انرژی کل نویز تقریباً برابر با انرژی کل سیگنال نویزی در واحدهای T-F پوشانده‌شده غالباً نویز[۳۳] (صفرها در IBM) است.

این مقاله از بخش‌های زیر تشکیل شده است. بخش ۲ توصیفی از سیستم پیشنهادی را برای تخمین SNR ورودی ارائه می‌دهد و جزئیات مراحل مختلف را در سیستم، موردبحث قرار می‌دهد. بخش ۳ به گزارش نتایج تجربی حاصل از ارزیابی سیستم پیشنهادی و مقایسه آن

با نتایج ارزیابی یک سیستم تخمین SNR قبلی (با استفاده از مراجع [۹] و [۱۰]) می‌پردازد. ملاحظات نهایی در بخش ۴ آمده است.

۲- فرمول‌بندی مسئله

سیستم پیشنهادی برای تخمین SNR ورودی در شکل ۱ نشان داده شده است. همان‌طور که در شکل مشاهده می‌شود، یک نسخه تخمین‌زده‌شده از IBM در فرآیند تخمین SNR به کار می‌رود. بااین‌وجود، ما در این مقاله سیستمی برای تخمین IBM به‌منظور جداسازی گفتار پیاده‌سازی نمی‌کنیم. به جای آن، فرض می‌کنیم سیگنال‌های پیش از ترکیب موجود هستند و بنابراین می‌توانیم از ماسک IBM واقعی استفاده کنیم (از این پس در همه جای این مقاله، عبارت ماسک IBM واقعی، به شکل خلاصه، IBM و یا ماسک IBM بیان می‌شود).

یک ماسک IBM با استفاده از یک SNR محلی به دست می‌آید. به‌طور خاص، به یک واحد زمانی-فرکانسی (T-F) که سیگنال هدف در آن غالب است، عدد یک و به غیر آن عدد صفر نسبت داده می‌شود [۱۱]. نسبت‌دادن اعداد یک و یا صفر را به واحدهای زمانی-فرکانسی، برچسب‌گذاری باینری واحدها[۳۴] می‌نامند. واضح است که هرگونه سیستم جداسازی گفتار مبتنی بر CASA که به تخمین IBM می‌پردازد، می‌تواند در ترکیب با سیستم پیشنهادی حاضر برای تخمین SNR به کار رود. این مطلب در شکل با خط‌چین نشان داده شده است.

۲-۱- محیط شنیداری[۳۵]

یک نمایش که به‌طور گسترده در سیستم‌های CASA پذیرفته شده است، نمایش زمانی-فرکانسی دو بعدی است. در این نمایش، بعد زمان از دنباله‌ای از فریم‌های زمانی و بعد فرکانس از بانکی از فیلترهای شنیداری (به‌عنوان‌مثال فیلترهای گاماتون[۳۶]) تشکیل شده است. سیگنال ورودی ابتدا، در حوزه فرکانس توسط بانکی متشکل از ۶۴ فیلتر گاماتون تجزیه می‌شود. فرکانس مرکزی این فیلترها بر روی مقیاس نرخ پهنای باند مستطیلی معادل[۳۷] (ERB) از ۵۰ هرتز تا ۸۰۰۰ هرتز توزیع شده است. پاسخ ضربه کانال فیلتر cام به‌صورت زیر است [۱۱]:

$$g_c(t)=t^3\exp(-2\pi b_c t)\cos(2\pi f_c t+\phi_c)u(t), \qquad (۱)$$

به‌طوری‌که b_c نرخ تنزل پاسخ ضربه[۳۸] مربوط به پهنای باند فیلتر، f_c فرکانس مرکزی فیلتر و ϕ_c فاز را نشان می‌دهد (ما در اینجا ϕ_c را مساوی با صفر تنظیم می‌کنیم).

در حوزه زمان، خروجی‌های هر کانال به فریم‌های زمانی ۲۰ میلی‌ثانیه‌ای با یک شیفت فریم ۱۰ میلی‌ثانیه‌ای تجزیه می‌شوند. نمایش T-F حاصل، کاکلی‌گرام[۳۹] نامیده می‌شود [۱۵]. در گام نهائی مدل محیطی، به‌منظور شبیه‌سازی نرخ شلیک[۴۰] عصب شنوایی، خروجی هر فیلتر گاماتون به شکل نیم‌موج[۴۱] یکسوسازی[۴۲] می‌شود. آثار اشباع[۴۳] با گرفتن ریشه دوم از سیگنال یکسوشده مدل می‌شود. به‌طوری‌که β افکتور فراکاهشی[۴۵] با مقدار پیش‌فرض ۱ است.

هرگاه m و c، به ترتیب، اندیس کانال و فریم را نشان دهند، پاسخ فیلتربانک در واحد $u(c,m)$ با $g(c,m)$ و پاسخ فیلتربانک پس از گذشتن از مراحل یکسوسازی و اشباع با $r(c,m)$ نمایش داده می‌شود [۱۱].

۲-۲- محاسبه انرژی

دامنه سیگنال گفتار در هر لحظه از زمان تغییر می‌کند. انرژی سیگنال ترکیب در هر واحد T-F به همراه برچسب باینری آن، نمایشی را فراهم می‌کند که سهم گفتار و نویز را منعکس می‌کند. انرژی کوتاه‌مدت[۴۴] در کانال c و فریم زمانی m به‌صورت رابطه (۲) تعریف می‌شود [۱۱]:

$$E(c,m)=10\log_{10}\left(\sum_{n=1}^{N}[r(c,mN/2-N/2+n)w(c,m)]^2\right), \qquad (۲)$$

به‌طوری‌که $r(c,m)$ پاسخ یکسوسازی‌شده و غیرخطی‌شده فیلتربانک گاماتون در واحد $u(c,m)$، N طول فریم و w پنجره مستطیلی است. روش دوم، محاسبه انرژی با استفاده از خروجی‌های فیلتربانک گاماتون، $g(c,m)$، به‌صورت رابطه (۳) است [۱۱]:

$$E(c,m)=10\log_{10}\left(\sum_{n=1}^{N}[g(c,mN/2-N/2+n)w(c,m)]^2\right). \qquad (۳)$$

در این مقاله، سیستم پیشنهادی با استفاده از هر دو روش محاسبه انرژی، پیاده‌سازی شده است.

۲-۳- تخمین **SNR** ورودی

فرض کنید c اندیس کانال، $E(c,m)$ انرژی سیگنال ترکیب در واحد $u(c,m)$ برحسب دسی‌بل و $y(c,m)$ برچسب باینری تخمین‌زده‌شده برای آن را نشان می‌دهد. ما دو کمیت را در هر کانال محاسبه می‌کنیم: کمیت نخست، انرژی کلی سیگنال گفتار است که در واحدهای فعال IBM (یعنی واحدهایی که در آن‌ها انرژی سیگنال گفتار بیش‌تر از انرژی سیگنال نویز است) به شکل زیر محاسبه می‌شود:

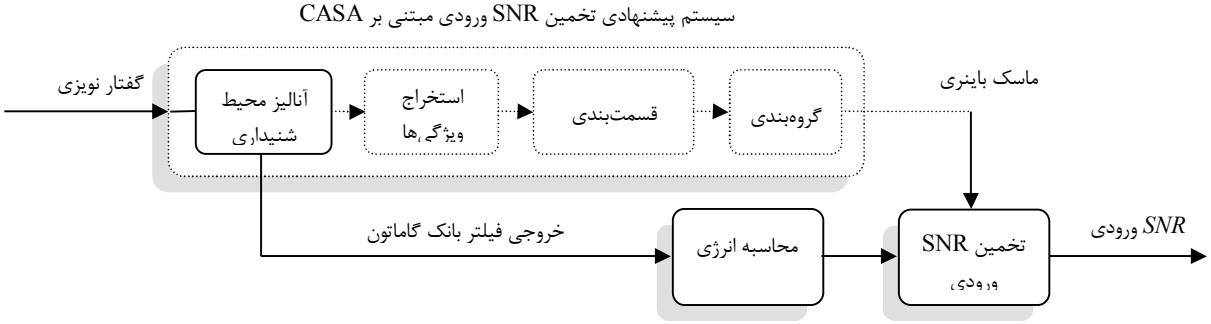
$$E_S(c)=\sum_m E(c,m).y(c,m). \qquad (۴)$$

کمیت دوم، انرژی کلی نویز است که در واحدهای غیرفعال (یعنی واحدهایی که در آن‌ها انرژی سیگنال نویز بیش‌تر از انرژی سیگنال گفتار است) به شکل رابطه (۵) محاسبه می‌شود:

$$E_N(c)=\sum_m E(c,m).(1-y(c,m)). \qquad (۵)$$

درنهایت، SNR ورودی سیگنال ترکیب، با استفاده از روابط (۴) و (۵)، توسط رابطه (۶) تخمین زده می‌شود:

$$SNR_{est}=10\times\log_{10}\left(\frac{\sum_c E_S(c)}{\beta\times\sum_c E_N(c)}\right), \qquad (۶)$$



شکل ۱: بلوک‌دیاگرام سیستم پیشنهادی تخمین **SNR** ورودی. گفتار نویزی ابتدا، توسط یک مدل محیط شنیداری آنالیز می‌شود. سپس، **IBM** طی مراحل متداول یک سیستم **CASA** تخمین زده می‌شود (نقطه‌چین). با فرض در دسترس‌بودن سیگنال‌ها پیش از ترکیب، **IBM** واقعی در فرآیند تخمین **SNR** به کار می‌رود.

۳- نتایج تجربی

سیستم پیشنهادی این مقاله بر روی ۷۰۰ سیگنال ترکیب آزموده شده و عملکرد آن مورد ارزیابی قرار گرفته است. این سیگنال‌ها از ترکیب ده جمله گفتار و ده سیگنال تداخل در هفت سطح SNR به دست آمده است. برای سیگنال‌های گفتار و تداخل، به ترتیب، از پایگاه داده‌های FARSDAT [۱۶] و Cooke [۱۷] استفاده شده است.

FARSDAT یک پایگاه داده گفتار فارسی پیوسته است که از ۶۰۰۰ گفته از ۳۰۰ گوینده با لهجه‌های مختلف فارسی موجود در ایران تشکیل شده است. این جملات از یک مجموعه ۳۹۰ جمله‌ای توسط گوینده‌های مختلف خوانده شده است. ده جمله از جملات فوق به شکل تصادفی انتخاب شده و در ترکیب با سیگنال‌های تداخل به کار رفته است.

سیگنال‌های تداخل Cooke به‌طور مشترک در مطالعات CASA به کار رفته‌اند و مقایسه ارزیابی‌های عینی[۴۶] الگوریتم‌های مختلف را آسان می‌کنند [۲۰-۱۸]. سیگنال‌های نویز در پهنای باند و ساختار زمانی تفاوت دارند و از ضبطی از یک آزمایشگاه آموزشی، موسیقی راک، صداهای باریک‌باند و پهن‌باند و جملات گفتاری دیگر تشکیل شده است.

فهرست دادگان مورداستفاده در ارزیابی سیستم پیشنهادی در جدول ۱ آمده است.

۳-۱- ارزیابی توسط ماسک **IBM**

یک سیگنال ترکیب، به‌عنوان ورودی یک سیستم CASA، توسط جمع یک سیگنال گفتار و یک سیگنال تداخل در یک سطح خاص از نسبت سیگنال به نویز (SNR) تولید می‌شود. سیگنال‌های تداخل بریده می‌شوند یا با خود پیوند داده می‌شوند تا با طول سیگنال گفتار متناظر مطابقت کنند. یک معیار محلی[۴۷] (*LC*) برابر با صفر دسی‌بل برای تولید IBM در همه شرایط SNR ورودی به کار گرفته می‌شود. شرایط شبیه‌سازی در جدول ۲ نشان داده شده است.

ابتدا، ما عملکرد سیستم مبتنی بر محاسبه انرژی توسط رابطه (۲) را ارزیابی می‌کنیم. نتایج ارزیابی‌ها در شکل ۲، برای میانگین ۱۰۰ سیگنال ترکیب در هفت سطح مختلف SNR، نشان داده شده است.

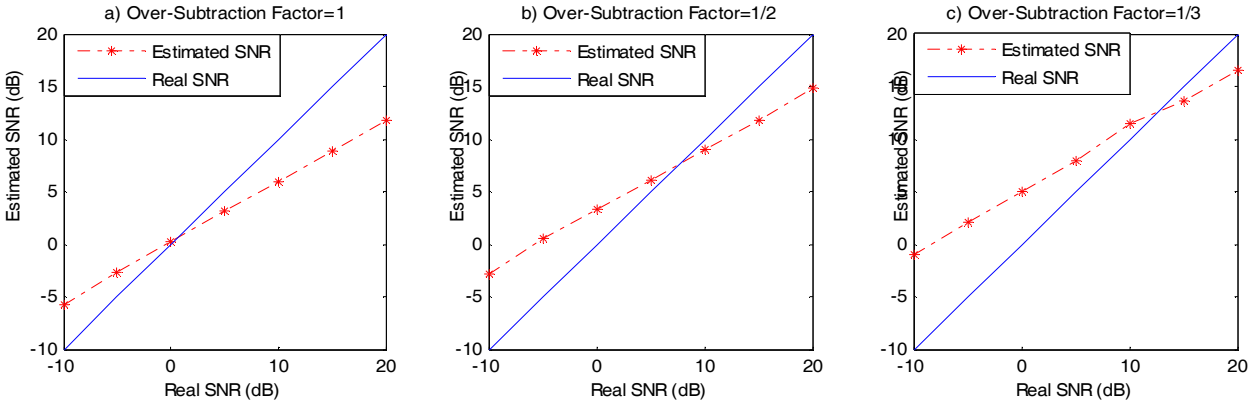
جدول ۱: دادگان مورداستفاده در شبیه‌سازی

شناسه نویز	محتوا	شناسه گفتار	گوینده	محتوا
N0	تون خالص 1 kHz	S04	مرد	اینجا مهد کشت و زرع است.
N1	نویز سفید	S12	مرد	بخل و بدخوئی از صفات ناپسند است.
N2	نویز انفجاری	S25	مرد	برادر نرگس صرع گرفت.
N3	نویز مهمانی	S33	مرد	این مطلب را در روزنامه درج کرد.
N4	موسیقی راک	S44	مرد	کتابت پاره شده است.
N5	سیگنال FM (آژیر)	S53	مرد	دیروز توت خوردم.
N6	زنگ تلفن	S61	زن	دیروز خیلی گرد و خاک خوردیم.
N7	گفته‌ای از پایگاه داده TIMIT بیان‌شده توسط گوینده زن	S72	زن	در داخل خلیج، ناو غرق شد.
N8	گفته‌ای از پایگاه داده TIMIT بیان‌شده توسط گوینده مرد	S81	زن	او با تبسم گفت: اینجا مأوای من است.
N9	گفته‌ای بیان‌شده توسط گوینده زن	S91	زن	دفتر صدبرگ پیدا شد.

جدول ۲: شرایط شبیه‌سازی

پارامتر	مقدار
فرکانس نمونه‌برداری (f_s)	۱۶۰۰۰ Hz
طول فریم (N)	۳۲۰ نمونه (۲۰ ms)
همپوشانی فریم‌ها ($N/2$)	۱۶۰ نمونه (۱۰ ms)
تعداد کانال‌ها	۶۴
محدوده فرکانسی	[۵۰، ۸۰۰۰] Hz
معیار محلی (LC)	۰ dB
فاکتور فراکاهشی (β)	۱، ۰/۵، ۰/۳۳[الف] ۱[ب]

الف- برای سیستم مبتنی بر محاسبه انرژی با استفاده از رابطه (۲)
ب- برای سیستم مبتنی بر محاسبه انرژی با استفاده از رابطه (۳)



شکل ۲: نتایج تخمین **SNR** ورودی مبتنی بر محاسبه انرژی با استفاده از رابطه (۲) با فاکتورهای فراکاهشی مساوی با الف) یک، ب) یک دوم و ج) یک سوم. میانگین نتایج در هر سطح **SNR** برای ۱۰۰ سیگنال ترکیب (در کل ۷۰۰ سیگنال ترکیب) به دست آمده است. نقاط تقاطع عبارت‌اند از: الف) (۰/۶، ۰/۶)، ب) (۷/۶، ۷/۶) و ج) (۱۲/۶، ۱۲/۶). به این ترتیب در هر شکل، مقادیری از **SNR** ورودی که به نقاط تقاطع یادشده نزدیک هستند، با دقت بالاتری تخمین زده می‌شوند.

همان‌طور که مشاهده می‌شود، شیب دو خط مبتنی بر مقادیر واقعی و تخمین‌زده‌شده SNR متفاوت است و در هر یک از قسمت‌های (الف) تا (ج) شکل، یک نقطه تقاطع وجود دارد. همچنین، مشاهده می‌شود که در این روش نقاط نزدیک به نقاط تقاطع بهتر تخمین زده می‌شوند. روی‌هم‌رفته، نزدیکی مقادیر تخمین‌زده‌شده به مقادیر واقعی نشان می‌دهد که الگوریتم مطابق انتظار کار می‌کند. مشاهده دیگر آن است که موقعیت نقاط تقاطع با تغییر فاکتور فراکاهشی در رابطه (۶) قابل تغییر است. بنابراین، در هر کاربرد می‌توان ناحیه کار (یعنی حوزه مقادیر SNR) را مشخص و فاکتور فراکاهشی را به مقدار مناسب تنظیم کرد تا نتیجه تخمین با خطای قابل‌قبول به دست آید. به‌علاوه، ما عملکرد سیستم مبتنی بر محاسبه انرژی با استفاده از رابطه (۳) را نیز ارزیابی کرده‌ایم.

میانگین نتایج برای ده سیگنال ترکیب در هفت سطح SNR، برای ده نوع نویز (در کل ۷۰۰ سیگنال ترکیب)، در شکل ۳ نشان داده شده است. در این حالت، به جز نویز N6 (نویز زنگ تلفن)، نتایج تخمین به مقادیر SNR واقعی بسیار نزدیک هستند.

۳-۲- ارزیابی توسط ماسک شبه **IBM**

همان‌طور که قبلاً ذکر شد، در این مقاله، سیستمی برای تخمین IBM و جداسازی سیگنال گفتار از سیگنال ترکیب پیاده‌سازی نشده است. با این وجود، در ارزیابی سیستم پیشنهادی، به‌منظور بررسی اثر خطاهای احتمالی در تخمین IBM، ماسکی به نام ماسک شبه IBM[۴۸]، به‌عنوان جایگزینی برای IBM تخمین‌زده‌شده توسط سیستم‌های CASA، به کار رفته است.

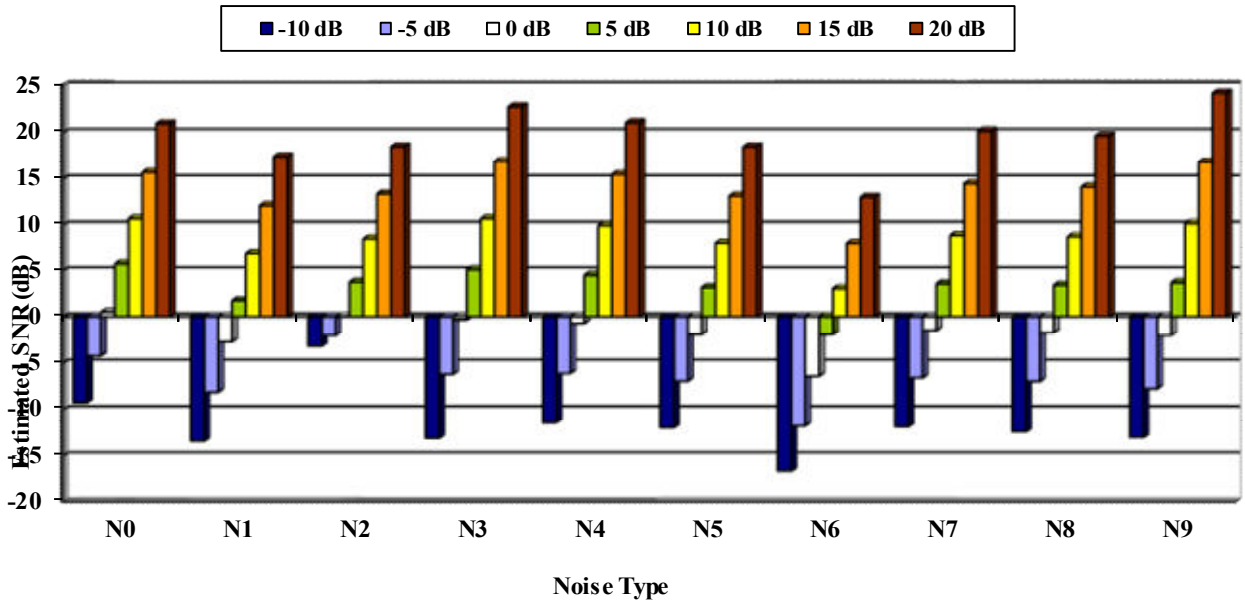
تولید ماسک شبه IBM برای هر سیگنال ترکیب شامل چهار مرحله است:

۱- تولید ماسک IBM برای آن ترکیب
۲- ایجاد یک ماسک با برچسب‌های تصادفی صفر و یک. این ماسک، ماسک خطا[۴۹] نامیده می‌شود.
۳- جمع باینری ماسک IBM و ماسک خطا. ماسک حاصل ماسک شبه IBM است.
۴- در پایان، میزان خطای ماسک شبه IBM نسبت به IBM محاسبه می‌شود.

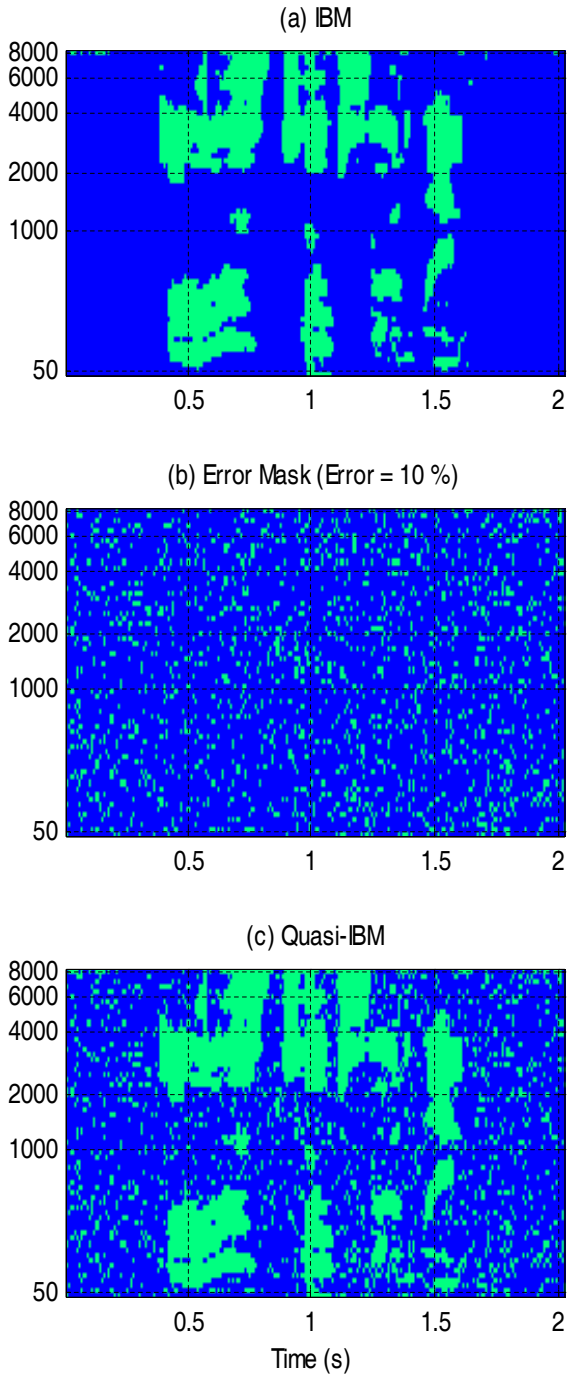
شکل ۴ ماسک‌های تولیدشده در هر یک از مراحل ۱ تا ۳ فوق را برای یک سیگنال ترکیب نشان می‌دهد. خطای محاسبه‌شده برای ماسک تولیدشده در این شکل ۱۰٪ است.

میانگین نتایج تخمین SNR برای ۱۰۰ ترکیب، در هفت موقعیت SNR با استفاده از ماسک‌های IBM و شبه IBM، بر مبنای محاسبه انرژی از طریق رابطه (۳)، در شکل ۵ نشان داده شده است. این نتایج (برای ماسک IBM) بسیار بهتر از تخمین SNR مبتنی بر محاسبه انرژی با استفاده از رابطه (۲) (شکل ۲) هستند، به‌علاوه، در این حالت نیازی به تعیین دامنه کاربرد سیستم برحسب مقادیر SNR ورودی و متعاقباً تنظیم فاکتور فراکاهشی نیز نیست. نزدیکی مقادیر تخمین‌زده‌شده و مقادیر واقعی نشان می‌دهد که الگوریتم پیشنهادی بر مبنای محاسبه انرژی با استفاده از خروجی فیلتربانک گاماتون (رابطه (۳)) در محدوده وسیعی از مقادیر SNR ورودی به درستی عمل می‌کند.

به‌علاوه، در شکل ۵، دو منحنی مربوط به مقادیر واقعی و تخمین‌زده‌شده SNR به کمک IBM، با اختلاف کمی از هم و با شیب‌های نسبتاً مشابه تا مقادیر SNR پایین امتداد یافته‌اند. این بدان معنی است که عملکرد سیستم پیشنهادی در شرایط SNR پایین کاهش نمی‌یابد. همچنین، نتایج برای تخمین SNR بر مبنای رابطه (۳) و با استفاده از ماسک شبه IBM قابل‌قبول ارزیابی می‌شود. توجه به این نکته حائز اهمیت است که استفاده از یک ماسک خطای تصادفی در هر سطح SNR منجر به تولید یک ماسک شبه IBM با میزان خطای متفاوت شده است. این خطا با کاهش سطح SNR ورودی افزایش می‌یابد. درنتیجه، می‌توان اثر خطاهای مختلف در تخمین IBM را بر عملکرد سیستم پیشنهادی ارزیابی نمود. به‌طورمثال، در شکل ۵، سیستم پیشنهادی در SNR ورودی ۲۰ دسی‌بل، با استفاده از ماسک‌های شبه IBM، با میانگین خطای حدود ۶/۳ درصد، SNR ورودی را با خطایی کم‌تر از ۲ دسی‌بل (اختلاف بین دو منحنی سبزرنگ (مقادیر SNR واقعی) و آبی‌رنگ (مقادیر SNR تخمین‌زده‌شده با ماسک شبه IBM)) تخمین می‌زند.



شکل ۳: نتایج تخمین **SNR** ورودی با روش مبتنی بر محاسبه انرژی با استفاده از رابطه (۳). ارزیابی‌ها مربوط به ۷۰۰ سیگنال ترکیب است و میانگین نتایج در ده نوع نویز مختلف و هفت سطح **SNR** نشان داده شده است. مقادیر واقعی **SNR** در بالای شکل برای هر نمودار میله‌ای نشان داده شده‌اند.



شکل ۴: ماسک‌های تولیدشده در فرآیند تولید یک ماسک شبه **IBM**. الف) ماسک **IBM** برای جداسازی سیگنال گفتار **S53** از یک ترکیب با **SNR=0 dB**، متشکل از گفتار **S53** و نویز مهمانی **N3**، ب) ماسک خطای تولیدشده توسط یک تابع تصادفی و ج) ماسک شبه **IBM** تولیدشده با خطای ۱۰٪ نسبت به ماسک **IBM** (حاصل‌جمع باینری ماسک‌های تولیدشده در قسمت‌های الف و ب).

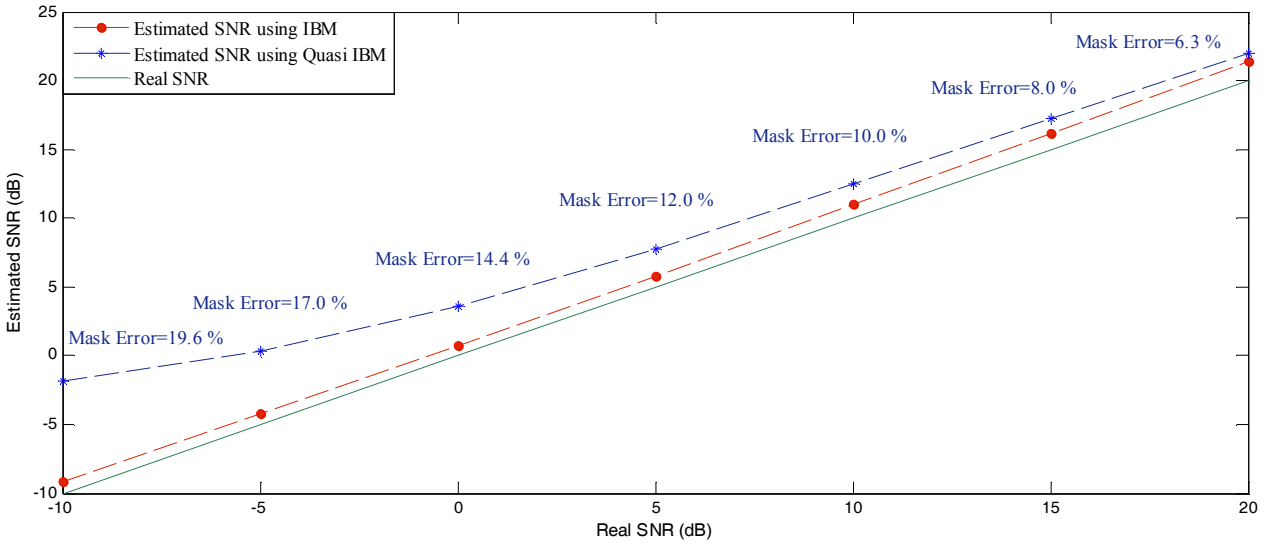
می‌توان نتیجه گرفت که سیستم پیشنهادی در مقابل خطاهای کم در تخمین ماسک IBM (مثلاً حدود ۶٪)، تا حد قابل‌قبولی پایدار است. اگرچه عملکرد سیستم با افزایش خطای ماسک به میزان قابل‌توجهی کاهش می‌یابد، اما ماسک‌های به‌دست‌آمده در مراحل میانی سیستم‌های CASA نیز، نسبت به ماسک‌های IBM خطای ناچیزی دارند [۲۱، ۲۲] و لذا، می‌توان نسبت به عملکرد مناسب سیستم پیشنهادی با استفاده از ماسک‌های به‌دست‌آمده در سیستم‌های CASA اطمینان داشت.

۳-۳- مقایسه عملکرد سیستم پیشنهادی با سیستم ردیابی **PSD** نویز مبتنی بر **MMSE**

در این روش، سیستم تخمین SNR بلندمدت معرفی‌شده در بخش مقدمه نام سیستم ردیابی PSD نویز مبتنی بر MMSE (مراجع [۹] و [۱۰]) نیز ارزیابی شده* و نتایج آن با نتایج سیستم پیشنهادی مقایسه شده است. در اینجا، برای اختصار از این سیستم با نام سیستم Hendriks یاد می‌شود. طول فریم و میزان انتقال فریم[۵۰]، به ترتیب، به مقادیر ۲۰ میلی‌ثانیه و ۱۰ میلی‌ثانیه تنظیم شده‌اند تا با فریم‌بندی سیستم پیشنهادی مطابقت داشته باشد. برای آنالیز زمانی- فرکانسی از بین‌های[۵۱]DFT به طول ۵۱۲ استفاده شده است. سایر پارامترها به مقادیر پیشنهادشده در مراجع [۹] و [۱۰] تنظیم شده است. توان نویز و اندازه مجذور گفتار با استفاده از الگوریتم‌های بهبود گفتار ارائه‌شده در مراجع [۹] و [۱۰] تخمین زده شده و با جمع این تخمین‌ها در زمان و فرکانس در حوزه DFT، SNR بلندمدت محاسبه می‌شود.

جدول ۳ مقادیر میانگین SNR تخمین‌زده‌شده و نیز قدرمطلق و انحراف‌معیار خطای تخمین را برای سیستم Hendriks و سیستم پیشنهادی در سطوح مختلف SNR ورودی نشان می‌دهد. مقادیر میانگین در هفت سطح SNR ورودی از ۱۰۰ سیگنال ترکیب به دست آمده است. سیستم پیشنهادی قادر به تخمین SNR با میانگین قدرمطلق خطای ۰/۹۵ دسی‌بل و میانگین انحراف‌معیار خطای ۲/۲۶ دسی‌بل است. این مقادیر برای سیستم Hendriks، به ترتیب، برابر با ۳/۷۷ دسی‌بل و ۱۰/۴۱ دسی‌بل است. همچنین، نتایج ارزیابی‌ها نشان می‌دهد، اگرچه در برخی موارد، به ویژه در شرایطی که سطح SNR ورودی بالاست، سیستم Hendriks، نسبت به سیستم پیشنهادی، تخمین بهتری از SNR سیگنال ترکیب ارائه می‌دهد، اما عملکرد سیستم پیشنهادی، برخلاف سیستم Hendriks، در سطوح SNR پایین کاهش نمی‌یابد.

* یک پیاده‌سازی از این الگوریتم در وبگاه به آدرس زیر:
http://siplab.tudelft.nl/content/mmse-based-noise-psd-tracking-algorithm (available on Nov.10.2014)
در دسترس می‌باشد که برای تولید نتایج گزارش‌شده در این مقاله به کار رفته است.



شکل ۵: نتایج تخمین **SNR** ورودی بر مبنای محاسبه انرژی با استفاده از خروجی فیلتربانک گاماتون (رابطه (۳)). نتایج ارزیابی برای میانگینی از ۱۰۰ ترکیب مختلف در هفت سطح **SNR** (در کل ۷۰۰ سیگنال ترکیب) نشان داده شده است. خط تو پُر سبز، نمودار مقادیر **SNR** واقعی، خط‌چین قرمز با نشانگر دایره، نمودار مقادیر **SNR** تخمین‌زده‌شده با استفاده از **IBM** و خط‌چین آبی با نشانگر ستاره، نمودار مقادیر **SNR** تخمین‌زده‌شده با استفاده از ماسک شبه **IBM** را نشان می‌دهند. میزان خطای ماسک شبه **IBM** در هر سطح **SNR** روی نمودار مربوطه درج شده است.

جدول ۳: میانگین مقادیر **SNR** تخمین‌زده‌شده و قدرمطلق و انحراف‌معیار خطای تخمین برای: (الف) سیستم **Hendriks** (**[۹] و [۱۰]**) و (ب) سیستم پیشنهادی (مبتنی بر ماسک **IBM** و محاسبه انرژی با استفاده از رابطه (۳)). مقادیر میانگین در هفت سطح **SNR** ورودی از ۱۰۰ سیگنال ترکیب به دست آمده است.

سیستم شبیه‌سازی‌شده	مقدار SNR ورودی (دسی‌بل)	۲۰	۱۵	۱۰	۵	۰	-۵	-۱۰	میانگین
(الف) سیستم **Hendriks**	SNR تخمین‌زده‌شده (دسی‌بل)	۲۰/۵۳	۱۶/۸۸	۱۲/۶۸	۸/۳۶	۴/۴۸	۱/۳۷	-۲/۸۸	۸/۷۷
	قدر مطلق خطای تخمین (دسی‌بل)	۰/۵۳	۱/۸۸	۲/۶۸	۳/۳۶	۴/۴۸	۶/۳۷	۷/۱۲	۳/۷۷
	انحراف‌معیار خطای تخمین (دسی‌بل)	۴/۸۸	۶/۵۰	۸/۴۰	۱۰/۴۱	۱۲/۵۱	۱۴/۳۸	۱۵/۸۰	۱۰/۴۱
(ب) سیستم پیشنهادی	SNR تخمین‌زده‌شده (دسی‌بل)	۲۱/۳۶	۱۶/۱۸	۱۰/۹۷	۵/۷۸	۰/۷۲	-۴/۲۵	-۹/۱۴	۵/۹۵
	قدر مطلق خطای تخمین (دسی‌بل)	۱/۳۶	۱/۱۸	۰/۹۷	۰/۷۸	۰/۷۲	۰/۷۵	۰/۸۶	۰/۹۵
	انحراف‌معیار خطای تخمین (دسی‌بل)	۲/۱۵	۲/۱۷	۲/۱۷	۲/۰۸	۲/۰۸	۲/۲۹	۲/۹۲	۲/۲۶

۴- نتیجه‌گیری

در این مقاله، روش جدیدی برای تخمین نسبت سیگنال به نویز (SNR) سیگنال ترکیب ارائه شده است که مبتنی بر روش آنالیز ترکیب شنیداری محاسباتی (CASA) است. ماسک IBM به‌عنوان هدف محاسباتی سیستم‌های CASA شناخته شده است. در طول مراحل مختلف میانی CASA (به‌عنوان‌مثال، قسمت‌بندی)، چندین تخمین اولیه از IBM به دست می‌آید. به‌علاوه، برخی سیستم‌های مهم CASA وجود دارند که طی مراحل مختلف پردازشی خود چندین بار برچسب‌گذاری واحدها را به اجرا درمی‌آورند. در چنین سیستم‌هایی، می‌توان ابتدا، از ماسک‌های اولیه‌ای که در مراحل میانی CASA به دست می‌آیند، برای تخمین SNR ورودی استفاده کرد. سپس، مراحل باقی‌مانده سیستم CASA برای تخمین ماسک IBM را بر اساس این تخمین SNR وفق داده و درنتیجه عملکرد سیستم را به میزان بیش‌تری بهبود بخشید.

مقدار SNR یک گفتار نویزی زمانی قابل تخمین است که هر دو جزء ترکیب، یعنی گفتار تمیز و نویز جمع‌شونده موجود باشند. هرگاه، تنها ترکیب این دو سیگنال در دسترس باشد، SNR گفتار باید تخمین زده شود. الگوریتم‌هایی وجود دارند که SNR ورودی را بدون سیگنال‌های مرجع تخمین می‌زنند. با این وجود، عملکرد این سیستم‌های تخمین SNR به شدت به عملکرد الگوریتم‌های آشکارسازی فعالیت صدا (VAD) وابسته است. در این مقاله، ما یک روش جدید برای تخمین SNR ورودی در سیستم‌های CASA پیشنهاد داده‌ایم. این روش نیازی به در دسترس‌بودن سیگنال گفتار تمیز یا بازسازی‌شده ندارد.

ارزیابی‌های اصولی و مقایسه‌ها با سیستم مبنایی Hendriks نشان می‌دهند که سیستم پیشنهادی در محدوده وسیعی از مقادیر SNR ورودی، شامل مقادیر SNR بسیار پایین، قادر به تخمین SNR ورودی، بر مبنای محاسبه انرژی با استفاده از خروجی فیلتربانک گاماتون، است. الگوریتم پیشنهادی، ازآنجاکه مبتنی بر سیستم CASA و ماسک باینری است، این قابلیت را دارد که در شرایط چالش‌انگیزی، مانند محیط‌های نویزی پژواک‌دار و شرایط حضور چندین‌گوینده[۵۲]، با جایگزینی ماسک باینری ایده‌آل به‌کاررفته در این پژوهش با ماسک‌های تخمین‌زده‌شده در شرایط پیچیده، مورداستفاده قرار گیرد.

سپاسگزاری

این پروژه تحت حمایت مرکز تحقیقات مخابرات ایران[۵۳] (ITRC) به انجام رسیده است. نویسندگان این مقاله از حمایت‌های مادی و معنوی این مرکز کمال سپاسگزاری را دارند.

مراجع

[1] M. Vondrasek, and P. Pollak, "Methods for speech SNR estimation: evaluation tool and analysis of VAD dependency," *Radioengineering*, vol. 14, no. 1, 2005.

[2] M. Kleinschmidt, and V. Hohmann, "Sub-band SNR estimation using auditory feature processing," *Speech Communication*, vol. 39, no. 1-2, pp. 47-64, 2003.

[3] H.G. Hirsch, "Estimation of noise spectrum and its applications to SNR-estimation and speech enhancement," *International Computer Science Institute, Berkeley, CA, Tech. Rep. TR-93-012*, 1993.

[4] R. Martin, "An efficient algorithm to estimate the instantaneous SNR of speech Signals," *Proc. Eurospeech*, pp. 1093-1096, 1993.

[5] Y. Ephraim, and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109-1121, 1984.

[6] *NIST Speech Quality Assurance (SPQA) Package v2.3*, 1994, Available online at: http://www.itl.nist.gov-/iad/mig/tools/ (available on Nov. 10, 2014).

[7] T.H. Dat, K. Takeda, and F. Itakura, "On-Line Gaussian mixture modeling in the log-power domain for signal-to-noise ratio estimation and speech enhancement," *Speech Commun.*, vol. 48, pp. 1515-1527, 2006.

[8] C. Kim, and R. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," *Proc. Interspeech*, pp. 2598-2601, 2008.

[9] R. Hendriks, R. Heusdens, and J. Jensen, "MMSE-based noise PSD tracking with low complexity," *Proc. IEEE ICASSP*, pp. 4266-4269, 2010.

[10] J. Erkelens, R. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741-1752, 2007.

[11] D.L.Wang, and G.J. Brown, *Computational Auditory Scene Analysis: Principles, algorithms and applications*, Ed., Hoboken, New York, IEEE Press, Wiley, 2006.

[12] G. Hu, and D.L. Wang, "Speech segregation based on pitch tracking and amplitude modulation," *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 79-82, 2001.

[13] N. Roman, D.L. Wang, and G.J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Am.*, vol. 114, no. 4, pp. 2236-2252, 2003.

[14] D.L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech Separation by Humans and Machines, P. Divenyi (Ed.), Kluwer Academic, Norwell, MA*, pp. 181-197, 2005.

[15] R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, ''An efficient auditory filterbank based on the gammatone function,'' *Applied Psychology Unit (APU)*, Report 2341 (Cambridge, UK), 1988.

[16] M. Bijankhan, J. Sheikhzadegan, M.R. Roohani, Y. Samareh, C. Lucas, and M. Tebiani, "The speech database of Farsi spoken language," *Proc. 5th Australian Int. Conf. on Speech Science and Technology (SST'94)*, pp. 826-831, 1994.

[17] http://www.dcs.shef.ac.uk/~martin/corpora/cookephd.tar.g-z, (available on Nov 10, 2014).

[18] D.P.W. Ellis, *Prediction-Driven Computational Auditory Scene Analysis*, Ph.D. Thesis, MIT, 1996.

[19] G. Hu, and D.L. Wang, "An auditory scene analysis approach to monaural speech segregation," *Topics in Acoustic Echo and Noise Control*, pp. 485-515, 2006.

[20] G. Hu, and D.L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks.*, vol. 15, no. 5, pp. 1135-1150, 2004.

[21] P. Dadvar, *Monaural Unvoiced Speech Segregation Based on Auditory Scene Analysis*, M.Sc. Thesis, Faculty of Electrical & Computer Eng., University of Tabriz, 2012.

[22] M. Geravanchizadeh, and P. Dadvar, "Monaural auditory-based unvoiced speech segregation using SNR-based subband spectral subtraction," *Acta Acustica United With Acustica*, vol. 100, no. 2, pp. 353-361, 2014.

زیرنویس‌ها

[1] Speech Enhancement
[2] Background Noise
[3] Signal-to-Noise Ratio (SNR)
[4] Compression
[5] Amplification
[6] Wiener Filter (WF)
[7] Spectral Subtraction
[8] Voice Activity Detection (VAD)
[9] Long-Time SNR
[10] Training Phase
[11] Short-Time SNR
[12] Low-Energy Envelope Tracking
[13] A priori SNR
[14] Decision Directed Approach
[15] Local SNR
[16] Speech Quality Assurance (SPQA)
[17] National Institute of Standards and Technology (NIST)
[18] Peak Signal-to-Noise Ratio
[19] Overestimate
[20] Expectation Maximization (EM)
[21] 2-Component Gaussian
[22] Learned Gaussians
[23] Bimodal Gaussian
[24] Gamma Distribution
[25] MMSE-Based Noise PSD Tracking
[26] Global Broadband SNR
[27] Computational Auditory Scene Analysis (CASA)
[28] Auditory Masking Phenomenon
[29] Ideal Binary Mask (IBM)
[30] Estimated IBM

[31] Time-Frequency (T-F) Units
[32] The Unmasked Speech Dominant T-F Units
[33] The Masked Speech Dominant T-F Units
[34] Binary Unit Labeling
[35] Auditory Periphery
[36] Gammatone Filters
[37] Equivalent Rectangular Bandwidth (ERB)
[38] Decay Rate of the Impulse Response
[39] Cochleagram
[40] Firing Rate
[41] Half-Wave
[42] Rectification
[43] Saturation Effects
[44] Short-Time Energy
[45] Over-Subtraction Factor
[46] Objective Evaluations
[47] Local Criterion (LC)
[48] Quasi-IBM
[49] Error Mask
[50] Frame Shift
[51] Bins
[52] Multi-Talker
[53] Iran Telecom Research Center (ITRC)