جداسازی تک‌گوشی گفتار صدادار مبتنی بر روش‌های جدید انتخاب واحدهای زمان- فرکانس در فرکانس‌های پایین و بالا

مسعود گراوانچی‌زاده[۱]، استادیار، صنم ایمانی شاملو[۲]، کارشناسی ارشد
۱- دانشکده مهندسی برق و کامپیوتر- دانشگاه تبریز- ایران- geravanchizadeh@tabrizu.ac.ir
۲- دانشکده مهندسی برق و کامپیوتر- دانشگاه تبریز- ایران- sanam.imani88@ms.tabrizu.ac.ir

چکیده: جداسازی گفتار در حضور نویز و سایر تداخل‌ها، یکی از مسائل چالش برانگیز به ویژه در مورد جداسازی تک‌گوشی می‌باشد. از آن جا که انتخاب دقیق واحدهای زمان-فرکانس تاثیر بسزایی در نتایج جداسازی دارد، در این مقاله، روش‌های جدیدی برای انتخاب واحدهای زمان-فرکانس در جداسازی تک‌گوشی گفتار صدادار ارائه می‌شود. در سیستم پیشنهادی ابتدا، واحدها با استفاده از همبستگی بین کانالی پوش پاسخ‌ها در مرحله‌ی جداسازی اولیه علامت‌گذاری می‌شوند. سپس سیستم در مرحله‌ی ردیابی گام و برچسب‌گذاری واحدها، ویژگی‌های متفاوتی را برای برچسب‌گذاری واحدها در فرکانس‌های بالا و پایین به کار می‌برد. در این مقاله از ویژگی تابع خودهمبستگی پوش بهبود یافته (EEACF) برای برچسب‌گذاری واحدهای زمان-فرکانس در حوزه فرکانس بالا استفاده شده است. تاثیر هر یک از این روش‌ها نیز در نتایج جداسازی مورد بررسی قرار گرفته است. ارزیابی سیستم جداسازی تک‌گوشی گفتار صدادار بر اساس معیارهای مختلف عینی برای انواع نویز مخلوط شده با سیگنال‌های گفتار انجام گرفته که نشان دهنده نتایج بهتر جداسازی نسبت به روش‌های معمول می‌باشد.

واژه‌های کلیدی: آنالیز ترکیب شنیداری محاسباتی، جداسازی گفتار صدادار، همبستگی بین کانالی پوش پاسخ، تابع خودهمبستگی پوش بهبود یافته.

Monaural Voiced Speech Segregation Based on New Methods of Choosing Time-Frequency Units in Low and High Frequencies

M. Geravanchizadeh[1], S. Imani Shamlou[2]

Faculty of Electrical & Computer Engineering, University of Tabriz, Tabriz, Iran

Abstract: Segregation of speech in presence of noise and other interferences is a challenging problem, especially in monaural case. Since accurate choice of time-frequency (T-F) units has great impact on segregation results, in this paper, we propose new methods for choosing T-F units in monaural voiced speech segregation. In the proposed system first, the units are marked using envelope cross-channel correlation in the initial segregation stage. Then, the system uses different features in the low and high frequencies for labeling of units in the pitch tracking and unit labeling stage. We use the enhanced envelope autocorrelation function (EEACF) for labeling T-F units in the high frequency range. We also consider the impact of each method on the segregation results. The evaluation of the monaural voiced speech segregation system based on different objective criteria is done for different types of noise mixed with speech signals which shows better segregation results than conventional methods.

Key Words: Computational Auditory Scene Analysis (CASA), Voiced Speech Segregation, Envelope Cross-Channel Correlation, Enhanced Envelope Autocorrelation Function (EEACF).

تاریخ ارسال مقاله: ۱۳۹۱/۷/۱۷
تاریخ اصلاح مقاله: ۸ /۱۳۹۲/۶
تاریخ پذیرش مقاله: ۱۳۹۲/۷/۹
نام نویسنده‌ی مسئول : دکتر مسعود گراوانچی‌زاده
نشانی نویسنده‌ی مسئول: ایران-تبریز-بلوار ۲۹ بهمن- دانشگاه تبریز- دانشکده مهندسی برق و کامپیوتر

۱- مقدمه

در حالی که جداسازی گفتار توسط ماشین به عنوان یک چالش بزرگ باقی‌مانده است سیستم شنیداری انسان توانایی قابل توجهی در این زمینه نشان می‌دهد و نیاز به یک سیستم کارآمد برای جداسازی گفتار مورد نظر در بسیاری از کاربردها محسوس است. به عنوان مثال، عملکرد سیستم تشخیص خودکار گفتار با وجود صداهای مزاحم به شدت کاهش می‌یابد. چنین سیستمی باید به یک جداکننده‌ی گفتار خوب مجهز شود. سیستم جداکننده‌ی گفتار همچنین در زمینه ارتباطات و مخابرات برای بهبود کیفیت گفتار و کاهش هزینه‌ی انتقال سیگنال غیرگفتار می‌تواند نقش موثری داشته باشد. علاوه بر این، صداهای مزاحم برای افرادی که دارای مشکل شنوایی هستند حتی وقتی از وسایل کمک شنوایی استفاده می‌کنند، یک مساله‌ی جدی هنگام گوش کردن به یک گوینده‌ی مشخص است. برای کمک به این افراد سیستم‌های کمک شنوایی باید طوری طراحی شوند که قابلیت جدا کردن عبارات مورد نظر را از ترکیب‌های صوتی داشته باشند.

در دنیای واقعی ترکیب گفتار و دیگر صوت‌ها توسط یک یا چند میکروفون، برای انجام پردازش گردآوری می‌شود. در محیط‌هایی نظیر مهمانی‌ها، استفاده از روش‌هایی مانند آنالیز مولفه‌های مستقل[۱] (ICA) [۱] یا فیلترینگ فضایی[۲] [۲] که در آن‌ها آرایه‌ای از میکروفون‌ها برای جداسازی گفتار موردنظر استفاده می‌شود نتایج مناسبی ندارند. همچنین، در برخی محیط‌ها استفاده از چندین میکروفون امکان‌پذیر نبوده و تنها انتخاب، حالت تک‌میکروفون خواهد بود. یک مثال از پردازش گفتار تک‌گوشی، تشخیص خودکار گفتار پخش‌کننده‌های رادیویی است. شنونده‌های انسانی توانایی قابل توجهی در جداسازی ترکیب شنیداری و توجه به یک صدای مشخص دارند. این روند ادراکی، آنالیز ترکیب شنیداری[۳] (ASA) نامیده می‌شود. بر همین اساس، آنالیز ترکیب شنیداری محاسباتی[۴] (CASA) زمینه‌ای مطالعاتی برای پیاده‌سازی ASA در ماشین‌هاست. در سال‌های اخیر، تحقیقات وسیعی در این زمینه صورت گرفته، از جمله Bregman کارهایی در این زمینه به چاپ رسانده است [۳]. Bregman فرآیند جداسازی شنیداری را، آنالیز ترکیب شنیداری (ASA) می‌نامد که در دو مرحله‌ی اصلی انجام می‌گیرد. مرحله‌ی اول، بخش‌بندی[۵] نامیده می‌شود که ترکیب شنیداری را به عناصر (یا بخش‌های) حسی تجزیه می‌کند به طوری که هر یک از آن‌ها مربوط به یک منبع یکتا باشد. مرحله‌ی دوم، گروه‌بندی[۶] نامیده شده و در آن بخش‌های مربوط به یک منبع در یک گروه قرار می‌گیرند. بخش‌بندی و گروه‌بندی توسط اصول ادراکی یا مشخصه‌های ASA صورت می‌گیرند.

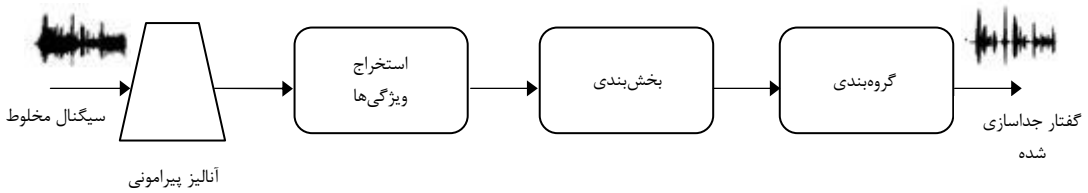
گفتار عادی شامل هر دو بخش صدادار و بی‌صدا می‌باشد. گفتار صدادار به بخش متناوب یا نیمه-متناوب سیگنال گفتار اشاره دارد و قسمت بزرگی از سیگنال صحبت را دربر می‌گیرد. تناوب و پیوستگی زمانی دو ویژگی اصلی مورد استفاده در جداسازی گفتار صدادار هستند. تناوب اشاره به این امر دارد که یک صوت متناوب، از گروهی از هارمونیک‌ها تشکیل شده است، به طوری که هر یک از آن‌ها یک مؤلفه‌ی فرکانسی بوده و فرکانس آن‌ها مضرب صحیحی از فرکانس پایه[۷] ($F0$) یا گام[۸] است. پیوستگی زمانی نیز اشاره به این مسئله دارد که سیگنال گفتار برای مدت معینی از زمان ادامه پیدا می‌کند و در طول این مدت، سیگنال معمولاً به طور هموار در زمان تغییر می‌کند. گرچه پیشرفت‌های قابل ملاحظه‌ای در زمینه‌ی جداسازی گفتار صدادار به دست آمده است، با این وجود عملکرد سیستم‌های جداسازی CASA در بخش‌های فرکانس بالا مناسب نیست. در کارهای اخیر انجام گرفته توسط Hu و Wang [۵، ۴] در جداسازی گفتار صدادار، از روش‌های متفاوتی برای جداسازی هارمونیک‌های جداشده[۹] و جدانشده[۱۰] استفاده شده است. با اینکه این سیستم‌ها نتایج بهتری نسبت به سیستم‌های پیشین [۷، ۶] ارائه می‌دهند ولی برای فرکانس‌های بالا (فرکانس‌های بالای یک کیلوهرتز) عملکرد قابل قبولی ندارند. مشکل اصلی این سیستم‌ها، نحوه‌ی برچسب‌گذاری واحدهای زمان-فرکانس[۱۱] است. گرچه بیشتر انرژی گفتار صدادار در بخش‌های فرکانس پایین متمرکز شده، با این وجود برچسب‌گذاری دقیق واحدها در هر دو ناحیه‌ی فرکانس پایین و فرکانس بالا می‌تواند تاثیر قابل توجهی در نتایج جداسازی داشته باشد.

با توجه به نکات مطرح شده، در این مقاله، روش‌های جدیدی برای برچسب‌گذاری واحدهای زمان-فرکانس پیشنهاد می‌شود. به بیان دقیق‌تر، در مرحله‌ی جداسازی اولیه، برای انتخاب واحدهای زمان-فرکانس، همبستگی بین کانالی پوش پاسخ در تمام فرکانس‌ها به کار رفته است. همچنین، در مرحله ردیابی گام و برچسب‌گذاری واحدها در فرکانس‌های بالا، از تابع خودهمبستگی پوش بهبودیافته[۱۲] (EEACF) برای حذف پیک‌های نادرست و یا مضارب پیک اصلی از منحنی تابع خودهمبستگی پوش[۱۳] (EACF)، استفاده شده است.

مقاله‌ی حاضر به صورت زیر سازمان یافته است. در بخش ۲، مراحل مختلف یک سیستم جداسازی گفتار متداول مبتنی بر آنالیز ترکیب شنیداری توضیح داده می‌شود. در بخش ۳، جزئیات سیستم پیشنهادی جهت جداسازی تک‌گوشی گفتار صدادار مورد بررسی قرار می‌گیرد. نتایج به دست آمده و مقایسه‌ها در بخش ۴ ارائه می‌شوند. در بخش ۵، نتیجه‌گیری کلی بیان می‌شود.

۲- پیش زمینه‌ای از سیستم‌های CASA

در این قسمت، پیش‌زمینه‌ای در مورد سیستم‌های جداسازی تک‌گوشی گفتار بر اساس CASA ارائه می‌شود که در مورد اکثر سیستم‌های CASA مشابه می‌باشد. بلوک دیاگرام کلی چنین سیستمی در شکل (۱) نشان داده شده است.



شکل (۱): بلوک‌دیاگرام کلی سیستم جداسازی تک‌گوشی گفتار بر اساس CASA [۴].

۱-۲- آنالیز پیرامونی (تجزیه‌ی زمان- فرکانسی)

در این مرحله، فیلترینگ گوش درونی، با تجزیه‌ی ورودی در حوزه‌ی فرکانس توسط گروهی از فیلترهای گاماتون شبیه‌سازی می‌شود. فیلتربانک گاماتون[۱۴] که از مشاهدات روان-آکوستیکی[۱۵] محیط شنیداری به دست آمده، مدلی استاندارد از گوش درونی است [۸].

ورودی توسط فیلتربانک گاماتون با ۱۲۸ کانال، با مراکز فرکانسی نیمه-لگاریتمی از 80 Hz تا 5000 Hz تجزیه می‌شود. پاسخ ضربه‌ی فیلتر گاماتون با فرکانس مرکزی f به صورت زیر می‌باشد:

$$g(f,t)=\begin{cases} b^{a}t^{a-1}e^{-2\pi bt}\cos(2\pi ft) & t\geq 0 \\ 0 & \text{else} \end{cases} \quad (۱)$$

که در آن $a=4$ مرتبه‌ی فیلتر بوده و b پهنای باند مستطیلی معادل[۱۶] (ERB) است. رابطه‌ی کلی بین ERB و فرکانس مرکزی f از رابطه زیر به دست می‌آید [۹]:

$$ERB(f)=0.108f+24.7\text{Hz} \quad (۲)$$

پهنای باند با افزایش فرکانس مرکزی افزایش می‌یابد. پاسخ هر فیلتر گاماتون توسط مدل سلول‌های مویی داخلی مدیس (Meddis) [۱۰] تبدیل می‌شود. این مدل ویژگی‌های شناخته شده‌ی سلول‌های مویی را، از قبیل یکسوسازی[۱۷]، اشباع[۱۸] و قفل‌شدگی فازی[۱۹]، شبیه‌سازی می‌کند. در هر کانال فیلتر، خروجی به فریم‌های زمانی 20 ms با هم‌پوشانی 10 ms بین فریم‌های متوالی تقسیم‌بندی می‌شود. با فیلترینگ میان‌گذر گاماتون و پنجره‌بندی زمانی، ورودی به نمایش زمان-فرکانس دوبعدی یا مجموعی از واحدهای زمان-فرکانس تجزیه می‌شود. نمایش دوبعدی به دست آمده حلزون‌نگاشت[۲۰] (کاکلیگرام) نامیده می‌شود.

۲-۲- استخراج ویژگی‌ها

در این مرحله، ویژگی‌های لازم برای مراحل بخش‌بندی و گروه‌بندی استخراج می‌شوند که شامل خودهمبستگی پاسخ فیلتر و خودهمبستگی پوش پاسخ فیلتر، همبستگی بین کانالی و همبستگی بین کانالی پوش‌ها و گام غالب در هر فریم می‌باشند.

زمانی که ترکیب ورودی شامل سیگنال متناوب است در فرکانس‌های بالا، برخی از کانال‌های فیلتر به چندین هارمونیک پاسخ می‌دهند. دامنه‌ی چنین پاسخ فیلتری مدوله شده و پوش پاسخ در فرکانس پایه‌ی، $F0$، سیگنال متناوب نوسان می‌کند. بنابراین، می‌توان از نرخ مدولاسیون دامنه، برای به دست آوردن $F0$ در ناحیه‌ی فرکانس بالا استفاده کرد. روش کلی برای به دست آوردن پوش پاسخ، یکسوسازی نیم-موج و سپس فیلترینگ پایین‌گذر است. از آن‌جا که هدف، استخراج نوسان‌های پوش مربوط به گام غالب است، در هر کانال از خروجی فیلتربانک، عمل فیلترینگ میان‌گذری که باند گذر آن محدوده‌ی مطلوب $F0$ گفتار مورد نظر است انجام می‌شود. بازه مطلوب یاد شده در این سیستم، حد [50 Hz, 550 Hz] در نظر گرفته شده است.

روش مناسب برای استخراج گام، استفاده از هم‌بستگی‌نگاشت[۲۱] (کوریلوگرام) است که خودهمبستگی متحرک[۲۲] هر یک از پاسخ فیلترها در طول فیلتربانک شنوایی است [۱۱]. هم‌بستگی‌نگاشت، نمایش شنوایی میان‌مرحله‌ای مناسبی بین محیط شنیداری و جداسازی فراهم می‌کند. برای واحد زمان-فرکانس u_{cm}، تابع خودهمبستگی[۲۳] (ACF) پاسخ (هم‌بستگی‌نگاشت) (A) و تابع خودهمبستگی پوش پاسخ (A_E)، به ترتیب، به صورت زیر به دست می‌آیند:

$$A(c,m,\tau)=\frac{\sum_n h(c,mT_m-nT_n)h(c,mT_m-nT_n-\tau T_n)}{\sqrt{\sum_n h^2(c,mT_m-nT_n)\sum_n h^2(c,mT_m-nT_n-\tau T_n)}} \quad (۳)$$

$$A_E(c,m,\tau)=\frac{\sum_n h_E(c,mT_m-nT_n)h_E(c,mT_m-nT_n-\tau T_n)}{\sqrt{\sum_n h_E{}^2(c,mT_m-nT_n)\sum_n h_E{}^2(c,mT_m-nT_n-\tau T_n)}} \quad (۴)$$

که روابط فوق، τ تاخیر، n زمان گسسته و $h(c,m)$ و $h_E(c,m)$ خروجی مدل سلول مویی و پوش آن در کانال c و فریم زمانی m هستند. شیفت زمانی از یک فریم تا فریم بعدی برابر با T_m بوده و T_n پریود نمونه‌برداری است. تناوب پاسخ فیلتر توسط پیک‌های تابع خودهمبستگی مشخص می‌شود. معادله‌ی (۳) حالت نرمالیزه شده‌ی ACF را محاسبه می‌کند.

همبستگی بین‌کانالی[۲۴]، تشابه میان پاسخ‌های دو کانال مجاور را اندازه‌گیری کرده و مشخص می‌کند که آیا پاسخ فیلترها مربوط به یک منبع یکسان هستند یا نه [۶،۱۲]. برای واحد زمان-فرکانس u_{cm}،

همبستگی بین کانالی (C) و همبستگی بین کانالی پوش[۲۵] (C_E) با واحد $u_{c+1,m}$، به ترتیب، به صورت زیر به دست می‌آیند:

$$C(c,m)=\frac{\sum_{\tau}[A(c,m,\tau)-\bar{A}(c,m)][A(c+1,m,\tau)-\bar{A}(c+1,m)]}{\sqrt{\sum_{\tau}[A(c,m,\tau)-\bar{A}(c,m)]^2\sum_{\tau}[A(c+1,m,\tau)-\bar{A}(c+1,m)]^2}} \quad (۵)$$

$$C_E(c,m)=\frac{\sum_{\tau}[A_E(c,m,\tau)-\bar{A}_E(c,m)][A_E(c+1,m,\tau)-\bar{A}_E(c+1,m)]}{\sqrt{\sum_{\tau}[A_E(c,m,\tau)-\bar{A}_E(c,m)]^2\sum_{\tau}[A_E(c+1,m,\tau)-\bar{A}_E(c+1,m)]^2}} \quad (۶)$$

که در آن‌ها $\bar{A}$ و $\bar{A}_E$، به ترتیب، میانگین A و A_E (تابع خودهمبستگی و تابع خودهمبستگی پوش) هستند.

در یک صوت متناوب، خودهمبستگی همه‌ی فیلترهای فعال در هم‌بستگی‌نگاشت پیکی را در تاخیر مربوط به تناوب نشان می‌دهد. بر اساس این ویژگی، روش هم‌بستگی‌نگاشت برای استخراج گام در تابع خودهمبستگی در تمام کانال‌ها استفاده می‌شود و یک پیک کلی در هم‌بستگی‌نگاشت مجموع[۲۶] مشخص می‌کند. اگر $s(m,\tau)$ هم‌بستگی‌نگاشت مجموع همه کانالها در فریم m-ام باشد در آن صورت داریم:

$$s(m,\tau)=\sum_{c}A(c,m,\tau) \quad (۷)$$

تناوب گام[۲۷] غالب در فریم m، یعنی $\tau_D(m)$، تأخیر مربوط به ماکزیمم $s(m,\tau)$ در محدوده‌ی مطلوب گام گفتار مورد نظر است. برای کانال‌هایی که گفتار مورد نظر در آن‌ها غالب است، پیک‌های خودهمبستگی آن‌ها با گام گفتار مورد نظر سازگار بوده و مجموع این خودهمبستگی‌ها پیک غالبی را در تناوب گام مربوطه نمایش می‌دهد.

۲-۳- بخش‌بندی

در این مرحله، واحدها بر اساس شرط‌های پیوستگی زمانی و همبستگی بین‌کانالی، به هم پیوسته و بخش‌ها را تشکیل می‌دهند. در کارهای انجام گرفته توسط Hu و Wang [۴،۵]، از همبستگی بین کانالی، C، برای علامت‌گذاری واحدهای قرار گرفته در فرکانس‌های پایین و همبستگی بین کانالی پوش پاسخ‌ها، C_E، برای علامت‌گذاری واحدهای قرار گرفته در فرکانس‌های بالا استفاده شده است.

۲-۴- گروه‌بندی

در مرحله‌ی گروه‌بندی، با مقایسه‌ی تناوب پاسخ واحدها با گام غالب در بخش‌های به دست آمده از مرحله‌ی قبل، رشته‌های پیش‌زمینه و پس‌زمینه‌ی اولیه تشکیل می‌شوند که به ترتیب مربوط به گفتار مورد نظر و تداخل هستند. سپس، به منظور به دست آوردن منحنی گام دقیق‌تر گام‌های جدید باز تخمین زده می‌شوند. در این مرحله، برچسب‌گذاری واحدها بر اساس منحنی گام به دست آمده انجام می‌شود. به این منظور، در مدل Hu و Wang [۴، ۵] از دو معیار زیر، به ترتیب، برای برچسب‌گذاری واحدهای فرکانس پایین (نامعادله (۸)) و برچسب‌گذاری واحدهای فرکانس بالا (نامعادله (۹)) استفاده شده است:

$$\frac{A(c,m,\tau_S(m))}{A(c,m,\tau_P(c,m))}>\theta_T \quad (۸)$$

$$\frac{A_E(c,m,\tau_S(m))}{A_E(c,m,\tau_P(c,m))}>\theta_A \quad (۹)$$

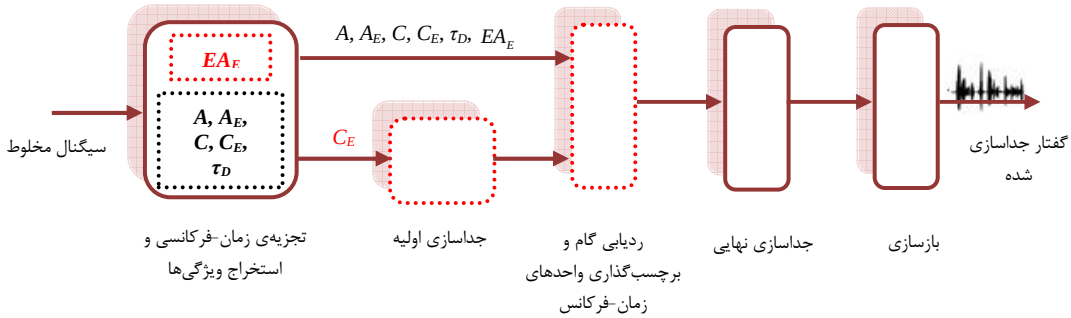
که در آن‌ها $A(c,m,\tau)$ و $A_E(c,m,\tau)$ به ترتیب تابع خودهمبستگی پاسخ و تابع خودهمبستگی پوش پاسخ، $\tau_S(m)$، مقدار گام به دست آمده از ردیابی گام و $\tau_P(c,m)$ تاخیر مربوط به مقدار ماکزیمم تابع خودهمبستگی مربوطه درمحدوده مطلوب گام است. مقادیر آستانه‌ی θ_T و θ_A، به ترتیب برابر با ۰/۸۵ و ۰/۷ در نظر گرفته شده‌اند. ادامه سیستم جداسازی در مدل Hu و Wang شامل گروه‌بندی نهایی رشته‌های مربوط به گفتار مورد نظر و نویز می‌باشد. در نهایت، رشته‌ی گفتار مورد نظر به دست آمده برای استخراج شکل موج نهایی بازسازی می‌شود.

۳- سیستم پیشنهادی جهت جداسازی گفتار صدادار

همانطور که قبلاً اشاره شد یکی از چالش‌های موجود در جداسازی گفتار صدادار، جداسازی واحدهای زمان-فرکانس گفتار مورد نظر از تداخل در محدوده‌ی فرکانسی بالا است. با توجه به مطالب قبل، در مدل Hu و Wang [۴،۵] از همبستگی بین کانالی، C، و همبستگی بین کانالی پوش پاسخ‌ها، C_E، برای علامت‌گذاری واحدهای زمان-فرکانس در مرحله بخش‌بندی استفاده شده است ولی این روش نتایج مناسبی به ویژه برای نویزهای هارمونیک ارائه نمی‌دهد.

بلوک دیاگرام سیستم پیشنهادی به منظور جداسازی گفتار صدادار در شکل (۲) نشان داده شده است. در این شکل بخش‌های مشخص شده با نقطه‌چین مربوط به نوآوری‌های سیستم پیشنهادی نسبت به سیستم جداسازی توسط Hu و Wang [۴،۵] است. در سیستم پیشنهادی، در مرحله‌ی بخش‌بندی اولیه، برای انتخاب واحدهای زمان-فرکانس، همبستگی بین کانالی پوش پاسخ‌ها، C_E، در تمام فرکانس‌ها به کار رفته است. به این ترتیب، واحدهای نویزی فرکانس پایین که در مدل Hu و Wang به اشتباه به عنوان واحد گفتار موردنظر انتخاب شده‌اند حذف می‌شوند. به عبارت دیگر، شرط انتخاب واحدها در مرحله‌ی بخش‌بندی اولیه به صورت $C_E(c,m)>\theta_C$ خواهد بود. سپس، برای ردیابی گام روش استفاده شده در مدل Hu و Wang به کار رفته است. همچنین، در سیستم پیشنهادی ویژگی دیگری با عنوان تابع خودهمبستگی پوش بهبود یافته (EEACF) استخراج می‌شود. از ویژگی EEACF برای حذف پیک‌های نادرست و یا مضارب پیک اصلی از منحنی تابع خودهمبستگی پوش (EACF)، در مرحله‌ی برچسب‌گذاری واحدها استفاده شده است. در واقع، تابع خودهمبستگی پوش بهبود یافته نمایش بهتری از **حداکثرهای اصلی تابع** خودهمبستگی پوش پاسخ ارائه می‌دهد. این راهکار، در برچسب‌گذاری

دقیق‌تر واحدهای فرکانس بالا که یکی از مشکلات جداسازی گفتار صدادار است تاثیر قابل ملاحظه‌ای می‌تواند داشته باشد. برای استخراج تابع خودهمبستگی پوش بهبود یافته از روش استفاده شده برای به دست آوردن تابع خودهمبستگی مجموع بهبود یافته[۲۸] برای ردیابی چندین گام[۲۹] در [۱۳] استفاده شده است.

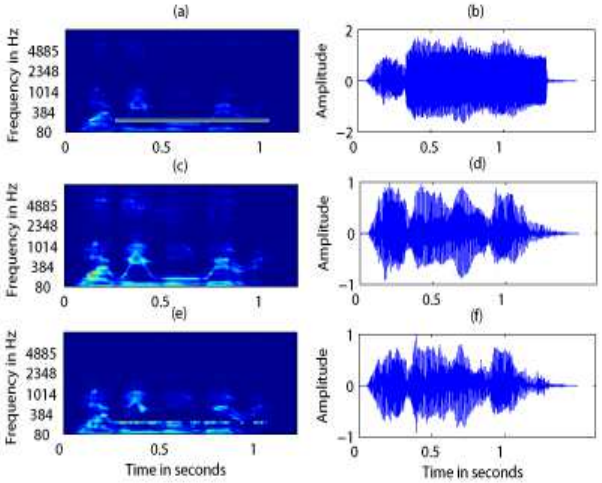


شکل (۲): بلوک‌دیاگرام سیستم پیشنهادی جداسازی تک‌گوشی گفتار صدادار بر اساس همبستگی بین کانالی پوش پاسخ‌ها (C_E) و تابع خودهمبستگی پوش بهبودیافته (EA_E)

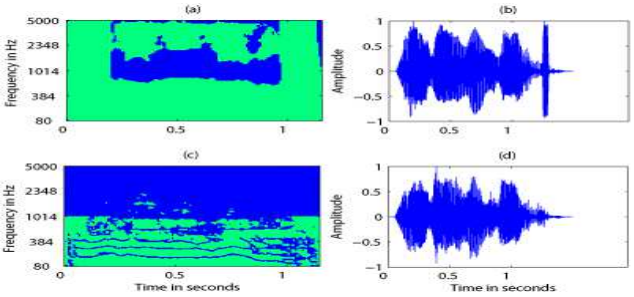
ابتدا، تابع خودهمبستگی پوش پاسخ در هر کانال برش داده می‌شود به طوری که مقادیر مثبت منحنی خودهمبستگی پوش پاسخ باقی می‌ماند. سپس، منحنی به دست آمده با ضریب دو در طول زمان گسترده شده و از سیگنال اصلی برش یافته‌ی تابع خودهمبستگی پوش پاسخ کم می‌شود. نتیجه‌ی به دست آمده بار دیگر برای حذف مقادیر منفی برش داده می‌شود. عملیات انجام گرفته باعث حذف حداکثرهای مربوط به تاخیرهای زمانی دو برابر زمان تاخیر اصلی از منحنی تابع خودهمبستگی پوش (EACF) می‌شود. همچنین، بخش‌های نزدیک به صفر منحنی تابع خودهمبستگی پوش حذف می‌شوند. در این عملیات، به منظور حذف مضارب بالاتر هر حداکثر، می‌توان منحنی تابع خودهمبستگی پوش را با ضریب N در طول زمان گسترده کرد و این مراحل را تا N بار تکرار کرد. در سیستم پیشنهادی، این مراحل به ازای $N = 1, 2, 3$ در طول هر یک از کانال‌ها برای تابع خودهمبستگی پوش پاسخ هر فریم انجام می‌گیرد. ویژگی به دست آمده EEACF با EA_E نمایش داده می‌شود. با استخراج این ویژگی، نامعادله (۱۰) جایگزین نامعادله (۹) در مدل Hu و Wang می‌شود:

$$\frac{EA_E(c,m,\tau_S(m))}{EA_E(c,m,\tau_P(c,m))} > \theta_A \qquad (۱۰)$$

در سیستم پیشنهادی، از مقادیر آستانه‌ی دیگری که در جدول (۱) آمده است، استفاده می‌شود. این مقادیر آستانه با آزمون و خطا به دست آمده و نتایج بهتری را در برچسب‌گذاری واحدها نسبت به مدل Hu و Wang ارائه می‌دهند. مرحله‌ی جداسازی نهایی در سیستم پیشنهادی، شامل تصحیح رشته‌های گفتار مورد نظر و نویز و گسترش بخش‌ها، همانند مدل Hu و Wang می‌باشد. خروجی این مرحله شامل تمام واحدهای زمان-فرکانس مربوط به گفتار مورد نظر می‌باشد که برای بازسازی صوت مورد نظر استفاده می‌شوند. در جدول (۲) ویژگی‌های به کار رفته در مدل Hu و Wang [۴، ۵] و سیستم پیشنهادی مقایسه شده‌اند. نتایج جداسازی سیستم پیشنهادی در شکل‌های (۳) و (۴) نشان داده شده‌اند.



شکل (۳): ردیف بالا سمت چپ: طیف‌نگاشت جمله‌ی "Why were you all weary?" با گوینده‌ی مرد مخلوط شده با نویز تون خالص. ردیف بالا سمت راست: شکل موج سیگنال مخلوط. ردیف وسط سمت چپ: طیف‌نگاشت جمله‌ی تمیز. ردیف وسط سمت راست: شکل موج جمله‌ی تمیز. ردیف پایین سمت چپ: طیف‌نگاشت جمله‌ی جداسازی شده توسط سیستم پیشنهادی. ردیف پایین سمت راست: شکل موج جمله‌ی جداسازی شده توسط سیستم پیشنهادی.



شکل (۴): ردیف بالا سمت چپ: رشته‌ی هدف (نواحی با رنگ روشن) شکل گرفته از تمام واحدهای زمان-فرکانس مربوط به گفتار مورد نظر (ماسک باینری ایده‌ال). ردیف بالا سمت راست: شکل موج بازسازی شده از ماسک باینری ایده‌ال. ردیف پایین سمت چپ: ماسک باینری ایده‌ال تخمینی توسط سیستم پیشنهادی. ردیف پایین سمت راست: شکل‌موج بازسازی شده از ماسک تخمینی خروجی.

جدول (۱): مقادیر عددی پارامترها

پارامترها	مقادیر عددی
T_m	۱۰ میلی‌ثانیه
T_n	۰/۰۶۲۵میلی‌ثانیه
τT_n	[۰ و ۱۲/۵] میلی‌ثانیه
θ_T	۰/۵۵
θ_A	۰/۴
θ_C	۰/۹۸۵

جدول (۲): مقایسه‌ی ویژگیهای به کار رفته در سیستم پیشنهادی و مدل Hu و Wang [۴، ۵].

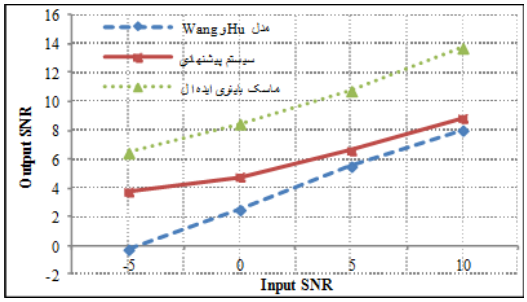
ویژگیهای مرحله‌ی برچسب‌گذاری واحدها		ویژگیهای مرحله‌ی جداسازی اولیه		
سیستم پیشنهادی	مدل Hu و Wang	سیستم پیشنهادی	مدل Hu و Wang	
(A)	(A)	(C_E)	(C)	فرکانس‌های پایین
(EA_E)	(A_E)	(C_E)	(C_E)	فرکانس‌های بالا

۴- نتایج شبیه‌سازی و مقایسه‌ها

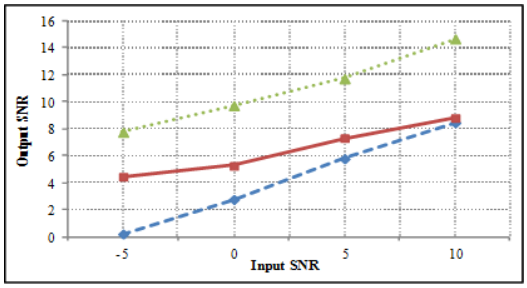
هر دو سیستم Hu و Wang [۴، ۵] و سیستم پیشنهادی توسط برنامه MATLAB شبیه‌سازی شده‌اند. در این شبیه‌سازی‌ها، جمله‌های گفتار و همچنین نویزها از دادگان Cooke [۱۴] انتخاب شده‌اند. سیگنال‌های گفتار تمیز شامل پنج جمله‌ی صدادار با گوینده‌ی مرد هستند. برای ایجاد سیگنال مخلوط ورودی، نویزهای مورد نظر با جمله‌ی گفتار تمیز در چهار سطح مختلف نسبت سیگنال به نویز[۳۰] (SNR) جمع شده‌اند. فرکانس نمونه‌برداری 16 KHz می‌باشد. نویزهای به کار رفته دارای تنوع قابل قبولی هستند که شامل: نویز تون خالص (N0)1-kHz، نویز رگبار (N2)، نویز مهمانی (N3)، نویز آژیر (N5)، نویز تلفن (N6) و نویز گفتار با گوینده‌ی زن (N7) می‌باشند. یکی از معیارهای ارزیابی عینی[۳۱] و سرراست، محاسبه‌ی SNR قبل و بعد از جداسازی، با استفاده از گفتار مورد نظر قبل از ترکیب با نویز است. این معیار با استفاده از رابطه زیر محاسبه می‌شود:

$$SNR = 10\log_{10}\left[\frac{\sum_n C^2(n)}{\sum_n (C(n)-S(n))^2}\right] \qquad (۱۱)$$

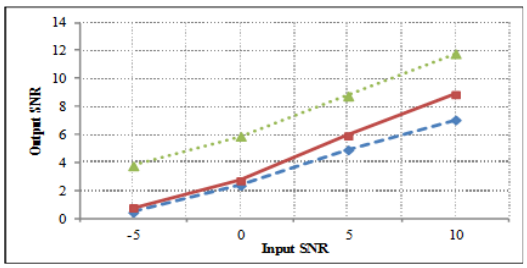
که در آن $C(n)$ بیانگر گفتار تمیز و $S(n)$ سیگنال گفتار جداسازی شده توسط سیستم جداسازی است. از آنجا که این معیار به طور معمول برای ارزیابی سیستم‌های جداسازی استفاده می‌شود، ابتدا، نتایج را بر این اساس مقایسه می‌کنیم که در شکل (۵) نشان داده شده است. مطابق این شکل، سیستم پیشنهادی بهبود قابل توجهی در SNR خروجی نسبت به مدل Hu و Wang



(الف)



(ب)



(ج)

شکل (۵): (الف) مقایسه‌ی مقدار میانگین نسبت سیگنال به نویز (SNR) سیگنال خروجی به دست آمده از مدل Hu و Wang [4, 5] و سیستم پیشنهادی برای تمام جملات مخلوط شده با تمام نویزها. (ب) مقایسه‌ی مقدار میانگین نسبت سیگنال به نویز (SNR) سیگنال خروجی به دست آمده از مدل Hu و Wang و سیستم پیشنهادی برای تمام جملات مخلوط شده با نویزهای غیر گفتار. (ج) مقایسه‌ی مقدار میانگین نسبت سیگنال به نویز (SNR) سیگنال خروجی به دست آمده از مدل Hu و Wang و سیستم پیشنهادی برای تمام جملات مخلوط شده با نویزهای گفتار و مهمانی.

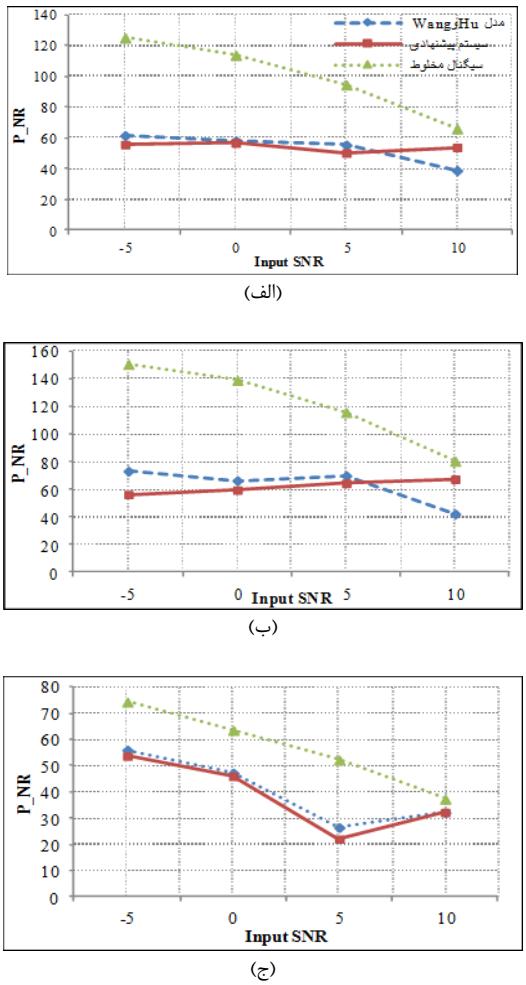
ارائه می‌دهد. از طرفی، میزان بهبودی در معیار SNR برای SNR-های ورودی پایین‌تر بیشتر است (شکل (۵)-الف). همچنین، SNR خروجی برای نویزهای گفتار و غیر گفتار نیز به طور جداگانه بررسی شده‌اند که در مورد نویزهای گفتار، نتایج به مقادیر مربوط به ماسک باینری ایده‌ال (IBM)[۳۲] نزدیک‌تر هستند (شکل (۵)-ج).

از آنجا که هدف محاسباتی در سیستم‌های CASA تخمین ماسک باینری ایده‌ال است عملکرد جداسازی را با مقایسه‌ی ماسک تخمینی و ماسک باینری ایده‌ال با دو معیار عینی دیگر نیز بررسی می‌کنیم: ۱) درصد اتلاف انرژی[۳۳] (P_{EL})، که میزان انرژی واحدهای زمان‌فرکانس هدف را که به عنوان تداخل برچسب‌گذاری شده‌اند نسبت به کل انرژی واحدهای هدف نشان می‌دهد، ۲) درصد نویز باقیمانده[۳۴] (P_{NR}) که میزان انرژی واحدهای زمان-فرکانس تداخل را که به عنوان هدف برچسب‌گذاری شده‌اند نسبت به کل انرژی واحدهای تداخل نشان می‌دهد. این معیارهای ارزیابی با استفاده از روابط زیر محاسبه می‌شوند:

$$P_{EL}=\frac{\sum_n e_1^2(n)}{\sum_n I^2(n)} \qquad (۱۲)$$

$$P_{NR}=\frac{\sum_n e_2^2(n)}{\sum_n O^2(n)} \qquad (۱۳)$$

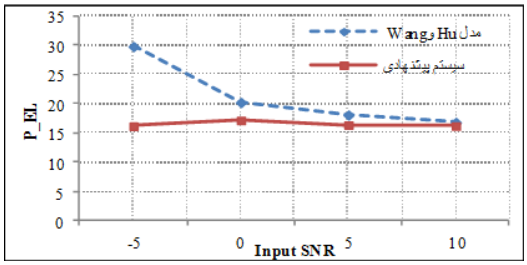
که در آن‌ها $O(n)$ سیگنال گفتار به دست آمده از سیستم جداسازی و $I(n)$ سیگنال گفتار بازسازی شده از ماسک باینری ایده‌ال است. $e_1(n)$ سیگنال موجود در $I(n)$ و حذف شده از $O(n)$ را نمایش می‌دهد و $e_2(n)$ بیانگر سیگنال موجود در $O(n)$ و حذف شده از $I(n)$ است. P_{EL} و P_{NR} معیارهای کاملی از خطا در سیستم جداسازی ارائه می‌دهند به طوری که یک سیستم قابل قبول باید مقادیر پایینی برای هر دو معیار داشته باشد. نتایج این معیارها در شکل‌های (۶) و (۷) نشان داده شده‌اند. گرچه در شکل (۷)-ب مقدار اتلاف انرژی سیستم پیشنهادی بیشتر از مدل Hu و Wang [۵، ۴] می‌باشد ولی بطور کلی می‌توان گفت که عملکرد سیستم پیشنهادی در دو معیار اخیر بهتر از مدل Hu و Wang [۵، ۴] است. با مقایسه بخش‌های (ب) و (ج) شکل‌های (۶) و (۷) می‌توان استنباط کرد که عملکرد سیستم پیشنهادی در مورد نویزهای گفتار و مهمانی بهتر از نویزهای غیر گفتار است. همچنین، اثر هر یک از ویژگی‌های به کار رفته در سیستم پیشنهادی شامل C_E و EA_E، در نتایج جداسازی مورد بررسی قرار گرفته است. به این منظور، سیستم پیشنهادی بدون استفاده از ویژگی EA_E و با در نظر گرفتن ویژگی A_E برای برچسب‌گذاری واحدهای فرکانس بالا پیاده‌سازی شده است. به عبارت دیگر، تمام واحدها در مرحله‌ی جداسازی اولیه با استفاده از همبستگی بین کانالی پوش پاسخ‌ها (C_E) علامت‌گذاری می‌شوند ولی برای برچسب‌گذاری واحدها در مرحله ردیابی گام و برچسب‌گذاری واحدها (مرحله‌ی سوم در شکل (۲)) از تابع خودهمبستگی پوش (A_E) (همانند مدل Hu و Wang) استفاده شده است. نتایج مقایسه در شکل (۸) نشان داده شده است. نتایج نشان می‌دهند که عدم استفاده از ویژگی EA_E برای برچسب‌گذاری واحدها در فرکانس بالا، تاثیر چندانی در معیار ارزیابی SNR خروجی نداشته (شکل (۸)-الف) و باعث بهتر شدن نتایج P_{EL} نسبت به سیستم پیشنهادی (شامل C_E و EA_E) می‌شود (شکل (۸)-ج) ولی این امر مقدار P_{NR} را افزایش می‌دهد (شکل (۸)-ب). علاوه بر این، برای مقایسه‌ی مستقیم شکل موج‌ها، SNR گفتار صدادار جداسازی شده را نسبت به مقدار به دست آمده‌ی آن از ماسک باینری ایده‌ال به صورت زیر محاسبه می‌کنیم:



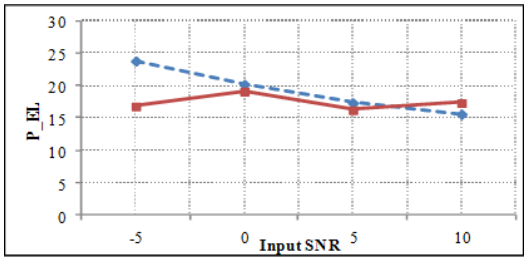
شکل (۶): (الف) مقایسه‌ی مقدار میانگین درصد نویز باقیمانده (P_{NR}) در سیگنال خروجی به دست آمده از مدل Hu و Wang [۴،۵] سیستم پیشنهادی برای تمام جملات مخلوط شده با تمام نویزها. (ب) مقایسه‌ی مقدار میانگین درصد نویز باقیمانده (P_{NR}) در سیگنال خروجی به دست آمده از مدل Hu و Wang و سیستم پیشنهادی برای تمام جملات مخلوط شده با نویزهای غیر گفتار. (ج) مقایسه‌ی مقدار میانگین درصد نویز باقیمانده (P_{NR}) در سیگنال خروجی به دست آمده از مدل Hu و Wang و سیستم پیشنهادی برای تمام جملات مخلوط شده با نویزهای گفتار و مهمانی.

$$SNR=10\log_{10}\left[\frac{\sum_n I^2(n)}{\sum_n (I(n)-O(n))^2}\right] \qquad (۱۴)$$

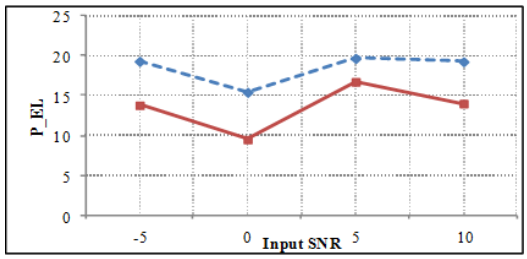
مقدار بالای این معیار به معنی شباهت بیشتر سیگنال خروجی به دست آمده از سیستم پیشنهادی به سیگنال به دست آمده از ماسک باینری ایده‌ال می‌باشد. نتایج SNR نسبت به ماسک باینری ایده‌ال و گفتار تمیز، برای نویزهای N0، N2، N5، N6 و N7 در پنج جمله‌ی صدادار و SNR ورودی 5- dB به ترتیب، در شکل‌های (۹) و (۱۰) نشان داده شده‌اند. همانطور که از شکل (۱۰) مشاهده می‌شود سیستم پیشنهادی در نویز N2، SNR خروجی بالاتری نسبت به ماسک باینری ایده‌ال دارد. به همین دلیل، نتیجه‌ی SNR در شکل (۹) برای این نویز



(الف)



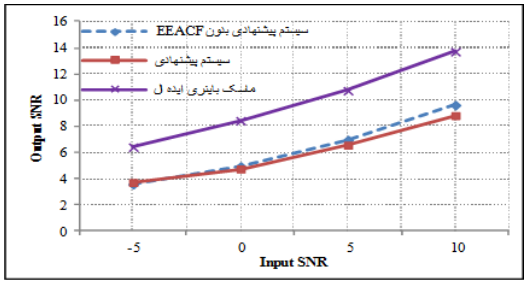
(ب)



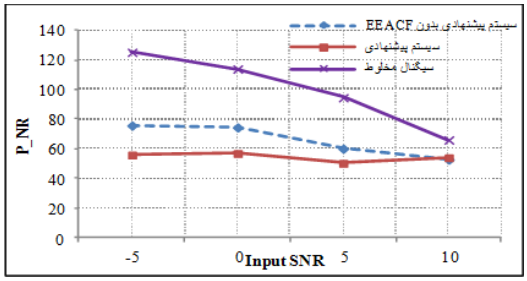
(ج)

شکل (۷): (الف) مقایسه‌ی مقدار میانگین درصد اتلاف انرژی (P_{EL}) در سیگنال خروجی به دست آمده از مدل Hu و Wang [۴۵] و سیستم پیشنهادی برای تمام جملات مخلوط شده با تمام نویزها. (ب) مقایسه‌ی مقدار میانگین درصد اتلاف انرژی (P_{EL}) در سیگنال خروجی به دست آمده از مدل Hu و Wang و سیستم پیشنهادی برای تمام جملات مخلوط شده با نویزهای غیر گفتار. (ج) مقایسه‌ی مقدار میانگین درصد اتلاف انرژی (P_{EL}) در سیگنال خروجی به دست آمده از مدل Hu و Wangو سیستم پیشنهادی برای تمام جملات مخلوط شده با نویزهای گفتار و مهمانی.

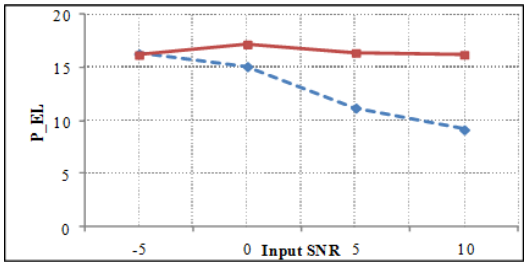
کمتر از مدل Hu و Wang می‌باشد. به عبارتی، رابطه (۱۴) معیار مناسبی برای بررسی عملکرد سیستم جداسازی نمی‌باشد. نتایج SNR خروجی (نسبت به گفتار تمیز) (رابطه (۱۱)) بدست آمده از مدل Hu و Wang، سیستم پیشنهادی و ماسک باینری ایده‌ال برای تمام جملات صدادار مخلوط شده با تمام نویزها در چهار سطح از SNR ورودی، در جدول (۳) نشان داده شده است. همانطور که ملاحظه می‌شود افزایش قابل توجهی در SNR خروجی سیستم پیشنهادی به ویژه در SNR-های ورودی پایین دیده می‌شود که نتیجه‌ی انتخاب دقیق واحدها در هر دو مرحله‌ی جداسازی اولیه و مرحله‌ی ردیابی گام و برچسب‌گذاری واحدهاست. در مورد SNR ورودی 10 dB، برای هر



(الف)



(ب)



(ج)

شکل (۸): مقایسه‌ی سیستم پیشنهادی بدون استفاده از EEACF (EA_E) در مرحله‌ی جداسازی اولیه و A_E در مرحله‌ی ردیابی گام و برچسب-گذاری واحدها) و سیستم پیشنهادی (شامل C_E و EA_E) برای تمام جملات مخلوط شده با تمام نویزها. (الف) میانگین نسبت سیگنال به نویز (SNR) سیگنال خروجی. (ب) میانگین درصد نویز باقیمانده (P_{NR}) در سیگنال خروجی. (ج) میانگین درصد اتلاف انرژی (P_{EL}) در سیگنال خروجی.

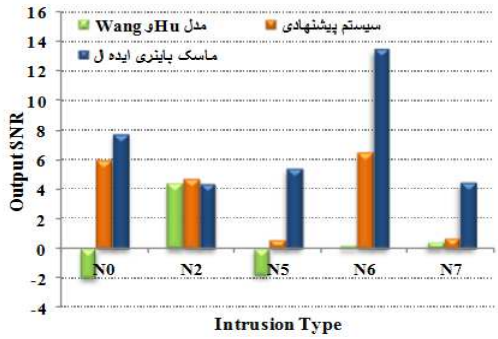
دو سیستم پیشنهادی و مدل Hu و Wang کاهش SNR خروجی مشاهده می‌شود که این مسئله می‌تواند به این دلیل باشد که در SNR-های ورودی بالا، سیگنال مورد نظر نسبت به نویز بسیار قوی‌تر بوده و عمل جداسازی در برخی موارد باعث از بین رفتن اطلاعات سیگنال مورد نظر می‌شود. همچنین، مقایسه‌ی نتایج جداسازی و ردیابی گام برای سیستم پیشنهادی و مدل Hu و Wang، به ترتیب، در شکل‌های (۱۱) و (۱۲) قابل مشاهده هستند. در شکل (۱۱) که

مربوط به نتایج جداسازی مخلوط جمله‌ی صدادار با نویز تون خالص (N0) به عنوان نویز هارمونیک است، طیف‌نگاشت و شکل‌موج جمله جداسازی شده توسط سیستم پیشنهادی، نسبت به مدل Hu و Wang، شباهت بیشتری به شکل‌های مربوطه‌ی جمله‌ی تمیز دارد.

جدول (۳): SNR خروجی به دست آمده از مدل Hu و Wang [۴،۵]، سیستم پیشنهادی و ماسک باینری ایده‌ال برای تمام جملات صدادار مخلوط شده با تمام نویزها در چهار سطح SNR ورودی.

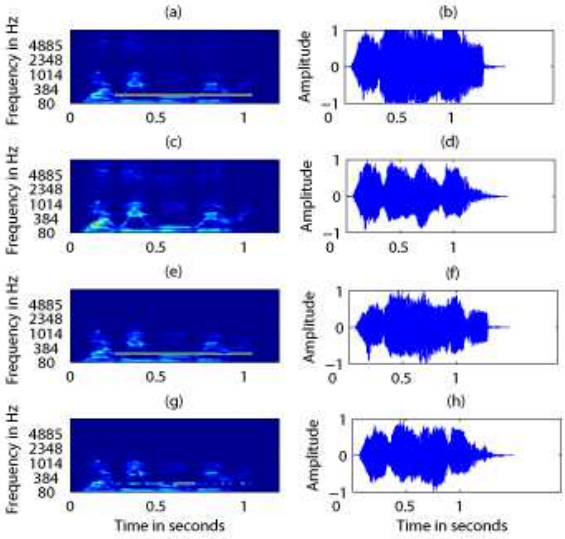
سیستم / SNR ورودی (dB)	مدل Hu و Wang	سیستم پیشنهادی	ماسک باینری ایده‌ال
-۵	۰/۲۲۵	۳/۷۲	۷/۱۱
۰	۲/۶۷	۴/۷۵	۹/۰۹
۵	۵/۴۹	۶/۸۵	۱۰/۴۷
۱۰	۸	۸/۸۴	۱۳/۲۷

شکل (۹): نتایج SNR نسبت به ماسک باینری ایده‌ال (مراجعه شود به فرمول (۱۴)) برای گفتارهای جداسازی شده. میله‌های نارنجی رنگ مربوط به نتایج سیستم پیشنهادی و میله‌های سبز رنگ نتایج حاصل از مدل Hu و Wang [۴،۵] می‌باشند. SNR مخلوط‌های ورودی dB 5- است.

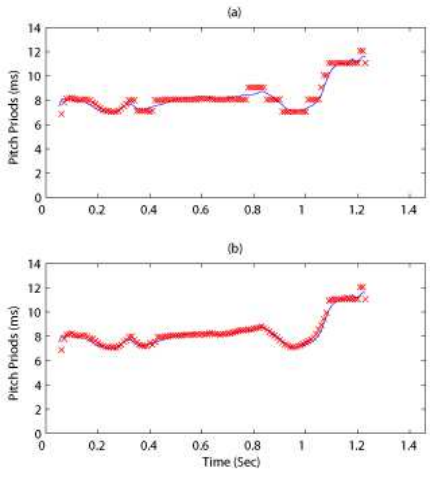


شکل (۱۰): نتایج SNR نسبت به گفتار تمیز (مراجعه شود به فرمول (۱۱)) برای گفتارهای جداسازی شده. میله‌های نارنجی رنگ مربوط به نتایج سیستم پیشنهادی، میله‌های سبز رنگ نتایج حاصل از مدل Hu و Wang [۴،۵] و میله‌های آبی رنگ مربوط به گفتار بازسازی شده از ماسک باینری ایده‌ال می‌باشند. SNR مخلوط‌های ورودی dB 5- است.

نتایج ردیابی گام در شکل (۱۲) نیز مقادیر گام دقیق‌تری برای سیستم پیشنهادی نسبت به مدل Hu و Wang نمایش می‌دهد که می‌توان آن را نتیجه‌ی علامت‌گذاری دقیق واحدها در مرحله‌ی جداسازی اولیه دانست. مقادیر واقعی گام جمله‌ی تمیز توسط نرم افزار[۱۶] Praat به دست آمده که با منحنی آبی رنگ در شکل نمایش داده شده است.



شکل (۱۱): ردیف اول از بالا سمت چپ: طیف‌نگاشت جمله‌ی "Why were you all weary?" با گوینده‌ی مرد مخلوط شده با نویز تون خالص. ردیف اول از بالا سمت راست: شکل موج سیگنال مخلوط. ردیف دوم از بالا سمت چپ: طیف‌نگاشت جمله‌ی تمیز. ردیف دوم از بالا سمت راست: شکل موج جمله‌ی تمیز. ردیف سوم از بالا سمت چپ: طیف‌نگاشت جمله‌ی جداسازی شده توسط مدل Hu و Wang [۴،۵]. ردیف سوم از بالا سمت راست: شکل موج جمله‌ی جداسازی شده توسط مدل Hu و Wang. ردیف چهارم از بالا سمت چپ: طیف‌نگاشت جمله‌ی جداسازی شده توسط سیستم پیشنهادی. ردیف چهارم از بالا سمت راست: شکل موج جمله‌ی جداسازی شده توسط سیستم پیشنهادی.



شکل (۱۲): نتایج ردیابی گام برای مخلوط جمله‌ی صدادار و نویز تون خالص. شکل بالا: مقادیر گام تخمینی توسط مدل Hu و Wang [۴،۵]. شکل پایین: مقادیر گام تخمینی توسط سیستم پیشنهادی. مقادیر گام تخمینی با نماد "x" نمایش داده شده‌اند. منحنی آبی رنگ منحنی گام به دست آمده از جمله‌ی تمیز توسط نرم افزار Praat می‌باشد.

۵- نتیجه

در این مقاله، سیستم جدیدی برای انتخاب واحدهای زمان-فرکانس در مرحله‌ی جداسازی اولیه و مرحله‌ی ردیابی گام و برچسب‌گذاری واحدها پیشنهاد شده است. بر اساس معیارهای ارزیابی عینی، افزایش قابل توجهی در SNR خروجی سیستم پیشنهادی نسبت به مدل Hu و Wang [۵، ۴] مشاهده می‌شود. هم چنین، نتایج دقیق ردیابی گام به دست آمده توسط سیستم پیشنهادی بیانگر انتخاب دقیق واحدها است. استفاده از همبستگی بین کانالی پوش پاسخ‌ها در مرحله‌ی جداسازی اولیه تاثیر مثبت زیادی در انتخاب اولیه‌ی واحدها دارد که ادامه عملکرد سیستم جداسازی، شامل ردیابی گام، را به طور قابل توجهی تحت تاثیر قرار می‌دهد. در سیستم پیشنهادی، برچسب‌گذاری واحدها در "مرحله‌ی ردیابی گام و برچسب‌گذاری واحدها" در فرکانس‌های بالا نیز برخلاف تابع خودهمبستگی پوش به کار رفته در مدل Hu و Wang، بر اساس تابع خودهمبستگی پوش بهبود یافته (EA_E) انجام گرفته است. کارهای آینده را می‌توان بر اساس معرفی ویژگی‌های جدید در مراحل بخش‌بندی و گروه‌بندی و استفاده از روش‌هایی جدید برای این مراحل ادامه داد.

سپاسگزاری

این پروژه تحت حمایت مرکز تحقیقات مخابرات ایران[۳۵] (ITRC) به انجام رسیده است. نویسندگان این مقاله از حمایت‌های مادی و معنوی این مرکز کمال سپاسگزاری را دارند.

مراجع

[1] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind Extraction of Dominant Target Sources Using ICA and Time-Frequency Masking," IEEE Transactions on Audio, Speech, Language Processing, Vol. 14, No. 6, pp. 2165-2173, 2006.

[2] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On Microphone-Array Beamforming from a MIMO Acoustic Signal Processing Perspective," IEEE Transactions on Audio, Speech, Language Processing, Vol. 15, No. 3, pp. 1053-1065, Mar. 2007.

[3] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, 1994.

[4] G. Hu and D. L. Wang, "An Auditory Scene Analysis Approach to Monaural Speech Segregation," in Topics in Acoustic Echo and Noise Control, pp. 485-515, Springer, 2006.

[5] G. Hu and D. L. Wang, "Monaural Speech Segregation Based on Pitch Tracking and Amplitude Modulation," IEEE Transactions on Neural Networks, Vol. 15, No. 5, pp. 1135-1150, 2004.

[6] D. L. Wang and G. J. Brown, "Separation of Speech from Interfering Sounds Based on Oscillatory Correlation," IEEE Transactions Neural Networks, Vol. 10, No. 3, pp. 684-697, May 1999.

[7] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, Upper Saddle River, NJ, USA, 2001.

[8] J. Holdsworth, I. Nimmo-Smith, R. D. Patterson, and P. Rice, "Implementing a Gammatone Filter Bank," MRC Applied Psych. Unit, 1988.

[9] B. R. Glasberg and B. C. J. Moore, "Derivation of Auditory Filter Shapes from Notched Noise Data," Hearing Research, Vol. 47, pp. 103-138, 1990.

[10] R. Meddis, "Simulation of Auditory-Neural Transduction: Further Studies," Journal of the Acoustical Society of American, Vol. 83, pp. 1056-1063, 1988.

[11] M. Slaney and R. F. Lyon, "On the Importance of Time- a Temporal Representation of Sound," in Visual Representations of Speech Signals, M. P. Cooke, S. Beet, and M. Crawford, Eds. New York: Wiley, pp. 95-116, 1993.

[12] G. J. Brown and M. P. Cooke, "Computational Auditory Scene Analysis," Computer Speech and Language, Vol. 8, pp. 297-336, 1994.

[13]T. Tolonen and M. Karjalainen, "A Computationally Efficient Multipitch Analysis Model," IEEE Transactions on Speech and Audio Processing, Vol. 8, pp. 708-716, Nov. 2000.

[14] M. P. Cooke, *Modeling Auditory Processing and Organization*, Cambridge University Press, 1993.

[15] D. L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Hoboken NJ: Wiley & IEEE Press, 2006.

[16] P. Boersma and D. Weenink, "Praat: Doing Phonetics by Computer," 2004.

[۱۷] صنم ایمانی شاملو، جداسازی تک گوشی سیگنال گفتار براساس ردیابی Pitch، پایان‌نامه کارشناسی ارشد، دانشگاه تبریز، تبریز، بهمن ۱۳۹۰

زیرنویس‌ها

[1] Independent Component Analysis (ICA)
[2] Spatial Filtering
[3] Auditory Scene Analysis (ASA)
[4] Computational Auditory Scene Analysis (CASA)
[5] Segmentation
[6] Grouping
[7] Fundamental Frequency
[8] Pitch
[9] Resolved Harmonics
[10] Unresolved Harmonics
[11] Time-Frequency (T-F) Units
[12] Enhanced Envelope Auto correlation Function (EEACF)
[13] Envelope Autocorrelation Function (EACF)
[14] Gammatone Filterbank
[15] Psychoacoustic Observations
[16] Equivalent Rectangular Bandwidth (ERB)
[17] Rectification
[18] Saturation
[19] Phase Locking
[20] Cochleagram
[21] Correlogram
[22] Running Autocorrelation
[23] Autocorrelation Function (ACF)
[24] Cross-channel Correlation

[25] Envelope Cross-channel Correlation
[26] Summary Correlogram
[27] Pitch Period
[28] Enhanced Summary Autocorrelation Function (ESACF)
[29] Multi-Pitch Tracking
[30] Signal-to-Noise Ratio (SNR)
[31] Objective Criterion
[32] Ideal Binary Mask (IBM)
[33] Percentage of Energy Loss (P_{EL})
[34] Percentage of Noise Residue (P_{NR})
[35] Iran Telecom Research Center (ITRC)