بررسی عوامل مؤثر در درک گفتار و ارائه یک روش ترکیبی برای افزایش میزان درک سیگنال گفتار نویزی

حمیدرضا ابوطالبی[i]

چکیده

این مقاله به بررسی موضوع افزایش قابلیت درک گفتار می‌پردازد. عوامل مؤثر در قابلیت درک گفتار در مواردی نظیر نویز، انعکاس، کم بودن پهنای باند و ساختار واجی گفتار، قابل دسته‌بندی است. با توجه به نقش برجسته‌تر نویز در این میان، به نویز به دید اصلی‌ترین عامل تخریب سیگنال نگریسته شده و روشی برای بهبود میزان فهم گفتار نویزی ارائه گردیده است. در مرحله اول، پس از تعیین واکداری/بی‌واکی هر قاب گفتار، فیلتر مناسب برای قاب مزبور معین و بر روی سیگنال نویزی اعمال می‌گردد. سپس با تعیین محل فورمنت‌ها و با استفاده از مدل تشدیدکننده درجه دوم برای هر فورمنت، فیلتری متناسب با طیف سیگنال طراحی و بر روی سیگنال خروجی مرحله قبل اعمال می‌شود. ارزیابی‌های انجام شده دلالت بر آن دارد که روش پیشنهادی در همه شرائط نویزی مورد آزمایش، منجر به بهبود کیفیت و افزایش قابلیت درک گفتار نویزی گردیده است. البته میزان این بهبود در مورد نویز سفید بیشتر و در مورد نویزهای پایین‌گذر کمتر است.

کلمات کلیدی

قابلیت درک گفتار، بهسازی گفتار، حذف نویز، فورمنت.

Studying on Speech Intelligibility and Presenting a Hybrid Method for Intelligibility Improvement of Noisy Speech

Hamid Reza Abutalebi

ABSTRACT

In this paper, the problem of speech intelligibility has been addressed. Effective parameters on speech intelligibility are categorized as: noise, reverberation, bandwidth, and phonemic structure. Due to the dominant role of noise in speech intelligibility, this research concentrates on the noise and introduces a hybrid method for improving the intelligibility of noisy speech. Firstly, for each frame of noisy input, a voiced/unvoiced decision is made by peak picking on the autocorrelation sequence. Based on this decision, appropriate noise shaping filter is applied on the input signal. As the next process, the formant frequencies are estimated and then, the frequency band around each formant is amplified. In this part, a sequential procedure has been employed to segment the frame spectrum into bands containing one formant. The performance of the proposed method has been evaluated through objective and subjective tests. These tests totally prove the capability of the proposed system in intelligibility improvement of noisy speech.

KEYWORDS

Speech intelligibility, Speech enhancement, Noise reduction, Formant.

[i] استادیار دانشکده مهندسی برق و کامپیوتر، دانشگاه یزد، habutalebi@yazduni.ac.ir.

۱- مقدمه

با رشد روز افزون استفاده از سیستم‌های ارتباط گفتاری در کاربردهای روزمره، نیاز به حفظ کیفیت گفتار به عنوان امری اجتناب ناپذیر مطرح گردیده است. شرائط ایده‌آل و عاری از نویزی که در کارها و شبیه‌سازی‌های آزمایشگاهی در نظر گرفته می‌شود، در بسیاری از کاربردهای واقعی، به طور جدی نقض گردیده و برقراری آنها زیر سئوال می‌رود. از این‌رو، مبحث بهسازی گفتار به عنوان یکی از ضرورت‌های کاربردی و عملی، از زمینه‌های فعال تحقیقاتی در سال‌های اخیر بوده است.

در قالب یک تعریف کلی، موضوع بهسازی گفتار[۱] عبارت است از تلاش برای بهبود عملکرد سیستم‌های ارتباط گفتاری در مواردی که سیگنال گفتار تحت تأثیر نویز، انعکاس، و یا سایر عوامل تخریبی واقع گردیده است [۱]،[٤].

کاربردهای گوناگون سیستم‌های ارتباط گفتاری (اعم از تلفن، رادیو، سمعک، سیستم بازشناسی گفتار و ...) و نیز تنوع عوامل تخریبی که موجب کاهش کیفیت و تخریب سیگنال گفتار می‌گردند، اهداف متفاوتی را برای فرایند بهسازی گفتار به دنبال آورده است. اگر میزان تخریب سیگنال گفتار زیاد نباشد، معمول‌ترین هدف از بکارگیری روش‌های بهسازی گفتار، بهبود کیفیت است. در کاربردهایی دیگر، افزایش دقت سیستم بازشناسی گفتار، کاهش خستگی شنونده و یا نظیر این موارد، هدف نهایی فرایند بهسازی گفتار است. این در حالی است که با افزایش میزان اثر عوامل تخریبی، کیفیت سیگنال گفتار خراب و خرابتر شده و کار به جایی می‌انجامد که افزایش قابلیتِ درک[۲] گفتار به عنوان هدف اصلی سیستم بهسازی گفتار مطرح می‌شود. با توجه به این مقدمه، برای بهبود قابلیت فهم گفتار دو رویکرد متصور است؛ رویکرد اول، استفاده از روش‌های معمول بهسازی گفتار در محیط‌های با قابلیت درک گفتاری کم، و رویکرد دوم، بکارگیری روش‌های خاص بهبود قابلیت درک گفتار است.

از دیدگاهی دیگر، روش‌های مطرح در زمینه بهسازی و بهبود قابلیت درک گفتار به دو دسته: ۱) روش‌های پیش‌پردازشی[۳]، و ۲) روش‌های پساپردازشی[٤] قابل تفکیک است [۱]. در روش‌های دسته اول، فرض بر آن است که سیگنال گفتار تمیز (قبل از تأثیر عوامل مخرب محیطی بر روی سیگنال) در دسترس بوده و با اعمال برخی پردازش‌ها سعی می‌گردد که بخش‌های مهم و ویژگی‌های مؤثر در بحث قابلیت فهم مورد تقویت و تأکید قرار گیرد تا پس از تخریب سیگنال توسط عوامل محیطی، امکان فهم گفتار وجود داشته باشد. روش‌های دسته دوم که دارای تعدد بیشتری نیز هستند، با فرض نبودن دسترسی به سیگنال گفتار تمیز طراحی شده‌اند. در این دسته، با کار بر روی سیگنال گفتار تخریب‌شده، سعی می‌شود اثر عوامل تخریبی تا جای ممکن جبران و پاکسازی گردد [٤].

روش‌های بهسازی گفتار تاکنون بیشتر به منظور بهبود کلی کیفیت گفتار و نه بطور خاص، برای افزایش قابلیت درک مطرح گردیده‌اند. از میان این روش‌ها، تفریق طیفی و فیلتر وینر از قدیمی‌ترین و پایه‌ای‌ترین‌ها هستند. به عنوان نمونه تکامل‌یافته فیلتر وینر می‌توان به روش‌های جدیدی نظیر β-Order MMSE-STSA[5] [٥] و یا OM-LSA[6] [٦] اشاره نمود. مبنای کار این دو، ارائه یک تخمین MMSE از سیگنال تمیز (به‌ترتیب، از توان β دامنه و از لگاریتم دامنه طیف) است. در روش OM-LSA، در هر قاب[۷] برای تخمین سیگنال گفتار تمیز، با وارد کردن احتمال حضور گفتار، سعی شده است که تخمین بهینه‌ای متناسب با ویژگی‌های نواحی سکوت و گفتار بدست آید.

در منابع و مراجع، معدود روش‌هایی نیز به منظور خاص بهبود قابلیت درک گفتار ارائه شده است. Niederjohn در [۷] از ترکیب یک فیلتر بالاگذر (با فرکانس قطع 1100 Hz و با شیب 12 dB/Octave) با یک فشرده‌ساز دامنه برای افزایش قابلیت فهم گفتار نویزی استفاده نمود. در تحقیقی دیگر [۸]، Niederjohn روشی پیش‌پردازشی ارائه داد که فورمنت[۸]های سیگنال تمیز را (قبل از ارسال و تخریب) استخراج و سپس تقویت نموده است. بدین طریق سیگنالی تولید شده است که از مقاومت بیشتری در مقابل عوامل کاهش‌دهنده قابلیت فهم برخوردار است. برای افزایش قابلیت فهم گفتار در محیط‌های با انعکاس زیاد نیز، Kitamura [۹] فیلترهای مدولاسیون را بکار گرفت که مبتنی بر ایده تقویت نواحی پرانرژی طیف گفتار است.

بسیاری از منابع و مراجع در بحث بهسازی و بهبود قابلیت درک گفتار، نویز را به عنوان اصلی‌ترین عامل تخریب سیگنال گفتار در نظر گرفته و به بررسی راهکارهای پاکسازی گفتار نویزی پرداخته‌اند. در این پژوهش نیز، اگرچه به بررسی عوامل مختلف مؤثر در کاهش قابلیت درک گفتار پرداخته شده، ولی از میان عوامل مختلف تخریب سیگنال گفتار، بحث بر روی نویز متمرکز می‌باشد.

بخش دوم این مقاله، به بررسی عوامل مؤثر در قابلیت درک گفتار پرداخته و از میان عوامل مورد بحث، نویز را به عنوان اصلی‌ترین عامل محدودکننده قابلیت درک مطرح می‌نماید. بخش سوم به معرفی جزئیات روش پیشنهادی برای بهبود قابلیت درک گفتار نویزی اختصاص دارد. در بخش چهارم، نحوه پیاده‌سازی روش تشریح گردیده و نتایج کمی و کیفی حاصل از سنجش کارآیی الگوریتم بیان می‌شود. بخش پایانی هم به

جمعبندی تحقیق می‌پردازد.

۲- عوامل مؤثر در قابلیت درک گفتار

بررسی‌های انجام شده حاکی از آن است که فرایند پیچیده فهم گفتار توسط عوامل زیادی مورد تأثیر قرار می‌گیرد که برخی از آنها منشأ فیزیکی داشته و بعضی نیز با ویژگی‌ها و ساختار زبان گفتار مرتبط است [۱۰]. ازمیان این عوامل، موارد زیر را می‌توان به عنوان مهمترین‌ها برشمرد.

۲-۱- نویز

اصلی‌ترین عامل در میزان قابلیت درک سیگنال گفتار، میزان و خواص طیفی (فرکانسی) نویز موجود در محیط است. در حالی که نسبت توان گفتار به توان نویز یا SNR می‌تواند به عنوان یک شاخص در تعیین میزان قابلیت درک گفتار مطرح گردد، لیکن طیف فرکانسی نویز و همپوشانی طیفی آن با طیف سیگنال گفتار، در کیفیت و قابلیت درک سیگنال گفتار اثر بسزایی دارد. در صورت همپوشان نبودن طیف سیگنال گفتار و طیف نویز، اگرچه از لحاظ شنیداری سیگنال نویزی مورد بحث، خستگی شنونده و آزردگی خاطر وی را بدنبال دارد ولی از لحاظ قابلیت فهم اثر چندانی نداشته و بعلاوه با یک فیلتر نمودن ساده می‌توان از اثر نویز موجود در محیط کاست.

۲-۲- انعکاس[۹]

میزان انعکاس موجود در سیگنال تا میزان زیادی به ابعاد محیط تولید گفتار بستگی دارد. از دیگر عوامل مؤثر در میزان انعکاس، نحوه جهت‌گیری میکروفون مورد استفاده و فاصله میکروفون تا گوینده است. با جهت‌گیری میکروفون به سمتی غیر از محل گوینده و یا با افزایش فاصله، نسبت دامنه مؤلفه‌های انعکاسی به مؤلفه اصلی افزوده شده و این پدیده، کاهش قابلیت درک گفتار را به همراه خواهد داشت [۱۱].

۲-۳- پهنای باند

در سیستم‌های مخابراتی، یکی از محدودیت‌های اثرگذار بر عملکرد سیستم، میزان پهنای باند است. در ارتباطات تلفنی که پهنای باند به محدوده ۳۰۰ هرتز تا ۳/۳ کیلو هرتز محدود می‌شود، بخشی از اطلاعات گفتاری، بویژه در نواحی بی‌واک از دست رفته و قابلیت درک گفتار را دچار اشکال می‌نماید [۱۲].

۲-٤- ساختار واجی[۱۰] گفتار

آنچه در واقع میزان قابلیت درک یک سیگنال گفتار را مشخص می‌نماید، برایند عوامل مؤثر در این مسأله است که در بحث فوق به مهم‌ترین آنها اشاره رفت. صرفنظر از اثر عوامل فیزیکی فوق، ساختار واجی سیگنال گفتار نیز اثر بسزایی در قابلیت درک ایفا می‌نماید. در بسیاری از زبان‌ها از جمله فارسی، انگلیسی، فرانسوی، آلمانی و روسی، بار معنایی و مفهومی عبارت به گونه‌ای قابل توجه در واج‌های همخوان قرار گرفته است [٤].

۳- معرفی روش پیشنهادی

بررسی و مقایسه اجمالی روش‌های بهسازی و بهبود قابلیت درک گفتار حاکی از آن است که اگرچه تمرکز کارهای گذشته بر روی بهسازی کلی گفتار بوده ولی در عین حال، روش‌های مطرح شده در این زمینه می‌تواند تا میزان مناسبی برای افزایش قابلیت درک گفتار نیز بکار رود. همچنین باید گفت روش مناسب برای هر کاربرد به شدت وابسته به نوع کاربرد و محیط مورد نظر بوده و نمی‌توان یک راه‌حل کلی برای بهسازی و بهبود قابلیت درک گفتار ارائه داد که برای همه کاربردهای گوناگون این بحث و نیز برای انواع عوامل تخریب سیگنال چاره‌ساز و کارآمد باشد [۱]. همان‌طور که پیشتر هم اشاره رفت، در این پژوهش توجه اصلی به نویز به عنوان مهم‌ترین عامل تخریب سیگنال گفتار معطوف گردیده و به بررسی و پیاده‌سازی روشی برای بهسازی و افزایش قابلیت درک گفتار نویزی اقدام شده است.

روش پیشنهادی این تحقیق، روشی پساپردازشی است که از تلفیق الگوریتم مطرح در [۲] و ایده تقویت فورمنت‌های سیگنال گفتار الهام گرفته شده است. پیشتر دیده شد که فورمنت‌های گفتار به عنوان باندهای فرکانسی مهم و پرانرژی سیگنال نقش تعیین‌کننده‌ای در قابلیت درک سیگنال گفتار دارا می‌باشند. در روش پیشنهادشده در اینجا، برای سیگنال گفتار نویزی، باندهای مهم فرکانسی (فورمنت‌ها) تعیین شده و پس از ترکیب اطلاعات بدست آمده با اطلاعات مربوط به واکداری/بی‌واکی، شکل‌دهی طیف نویز (تقویت منطقه باند مهم فرکانسی و تضعیف سایر مناطق فرکانسی) صورت می‌پذیرد.

۳-۱- تعیین واکداری/بی‌واکی برای سیگنال گفتار نویزی

واج‌های موجود در هر زبان به دو دسته واکدار و بی‌واک قابل تفکیک هستند. از لحاظ طیفی، واکدارها دارای تعدادی قله در فرکانسهای پایین (تا حدود ۱/۵ کیلوهرتز) می‌باشند که این قله‌ها در واقع هارمونیک‌های فرکانس گام است. در نقطه مقابل، بی‌واک‌ها معمولاً در فرکانس‌های پایین محتوای انرژی زیادی نداشته و تمرکز انرژی آنها در فرکانس‌های بالای ۳ کیلوهرتز است [٤].

بر اساس اطلاعات فوق، روش‌های مختلفی برای تعیین

واکداری یا بی‌واکی هر قاب گفتار می‌توان ارائه داد. بدیهی‌ترین روش، استفاده از طیف سیگنال و پردازش بر روی آن است. در صورت وجود حداقل یک قله فرکانسی در محدوده صفر تا ۱/۵ کیلوهرتز (با دامنه بلندتر از یک حد آستانه δ)، قاب مزبور واکدار در نظر گرفته شده و محل قله‌ها ثبت می‌شود. در غیر این صورت، قاب مورد نظر به عنوان بی‌واک تلقی خواهد شد. حد آستانه δ پارامتری است که رابطه مستقیم با توان متوسط سیگنال گفتار داشته و به صورت تجربی تعیین می‌گردد [۲].

در سیگنال‌های نویزی، وجود نویز قله‌های طیف واکدارها را دچار تحت تأثیر قرار داده و از دقت روش فوق در تعیین واکداری/بی‌واکی و محل قله‌ها (در صورت واکداری) می‌کاهد. راه حلی که در این مورد به نظر می‌رسد پردازش در حوزه زمان و بر روی تابع خودهمبستگی سیگنال زمانی است.

با توجه به این که دوره تناوب گام انسان در محدوده‌ای بین۳ تا ۱۵ میلی‌ثانیه تغییر می‌نماید، با بررسی دامنه تابع خودهمبستگی در اندیس‌های بین 0.003 Fs تا 0.015 Fs (که Fs فرکانس نمونه‌برداری است) و به شرط وجود قله‌ای بزرگتر از یک حد آستانه δ، قاب مزبور واکدار و محل قله (I_0) به عنوان دوره تناوب گام در نظر گرفته می‌شود. در صورت نبودن قله بلندتر از δ در تابع خودهمبستگی در محدوده زمانی یادشده، قاب مورد بحث بی‌واک فرض می‌گردد.

۳-۲- تعیین محل فورمنت‌ها

فورمنت‌ها، قطب‌های تابع تبدیل لوله صوتی و یا به تعبیر دیگر، صفرهای تابع تبدیل معکوس در آنالیز پیش‌بینی خطی (یا LPC[11]) می‌باشند. اگرچه استفاده از آنالیز LPC به عنوان یکی از اولین راه‌های تعیین فورمنت‌ها مطرح می‌باشد، و لیکن عملکرد آنالیز LPC در مدلسازی سیگنال نویزی ضعیف بوده و نتایج قابل قبولی را به دنبال ندارد.

در مدلسازی تولید گفتار و نیز در روش‌های سنتز قاعده‌مند گفتار (نظیر روش Klatt [۱۳])، اثر فورمنت‌ها به شکل مجموعه‌ای از تشدیدکننده‌های سری یا موازی در نظر گرفته می‌شود. در روش پیشنهادی، از مدل تشدیدکننده‌های سری استفاده شده و معیاری مبتنی بر تئوری فیلتر وینر برای تعیین مرز محدوده هر فورمنت بکار گرفته می‌شود.

۳-۲-۱ مدلسازی فورمنت‌ها

با این فرض که طیف سیگنال گفتار به گونه‌ای تقطیع شده که در هر باند فرکانسی تنها یک فورمنت وجود داشته باشد، هر فورمنت توسط یک تشدیدکننده با تابع تبدیل:

$$H_k(z) = 1/A_k(z) = 1/(a_{k,0} + a_{k,1}z^{-1} + a_{k,2}z^{-2}) \qquad (۱)$$

مدل می‌شود که توسط نویز سفید (به عنوان ورودی) تحریک گردیده است. به طور معمول، $a_{k,0} = 1$ فرض می‌شود. با توجه به تعریف فوق می‌توان گفت که فیلتر $A_k(z)$ فیلتر سفیدساز فورمنت و یا به بیانی، یک فیلتر میان‌نگذر[12] است که فرکانس حذفی آن همان فرکانس فورمنت می‌باشد. رابطه پارامترهای فیلتر $A_k(z)$ با فرکانس (F_k) و پهنای باند فورمنت مربوطه (B_k) به صورت رابطه (۲) می‌باشد:

$$a_{k,1} = \exp(-2\pi B_k), a_{k,2} = -2\exp(-\pi B_k)\cdot\cos(2\pi F_k) \qquad (۲)$$

می‌توان نشان داد [۱٤] که مقادیر بهینه ضرائب فیلتر $A_k(z)$ توسط فرمول‌های (۳) بدست می‌آید:

$$a_{k,1} = \frac{r_k(0)r_k(1) - r_k(1)r_k(2)}{r_k^2(0) - r_k^2(1)} \qquad (۳\text{-الف})$$

$$a_{k,2} = \frac{r_k(0)r_k(2) - r_k^2(1)}{r_k^2(0) - r_k^2(1)} \qquad (۳\text{-ب})$$

که در آن $r_k(m)$ مقدار تابع خودهمبستگی در باند k و در اندیس m بوده و با توجه به رابطه:

$$r_k(m) = \frac{1}{I}\sum_{i=i_{k-1}+1}^{i_k} |S(i)|^2 \cos(\frac{2\pi mi}{2I}) \qquad (٤)$$

از روی نمونه‌های FFT قاب مربوطه از سیگنال گفتار (یعنی $S(i)$ها) بدست می‌آید. در این رابطه، i_k مرز بالایی قطعه k-اُم از طیف سیگنال گفتار در قاب مورد پردازش و $2I$ طول FFT استفاده شده است. در ادامه در مورد نحوه تعیین مرزهای قطعات طیف گفتار بحث خواهد شد.

۳-۲-۲ چگونگی تشخیص مرز باندها

در قسمت قبل، چگونگی مدلسازی یک قطعه (باند) از طیف سیگنال گفتار به صورت یک مدل درجه دو بررسی شد. در این قسمت بر اساس تئوری فیلتر وینر، معیاری برای تعیین مرز هر باند حاوی یک فورمنت ارائه می‌شود. با فرض آن که فیلتر $A_k(z)$ یک فیلتر میان‌نگذر است که از سیگنال نویزی بخش سیگنال (فورمنت) را حذف می‌نماید، می‌توان مسأله حاضر را به شکل یک فیلتر وینر تقریبی در نظر گرفت که بر روی ورودی سیگنال نویزی (یعنی $x(n) = s(n) + n(n)$) عمل نموده و سعی در تولید خروجی $d(n) = n(n)$ دارد. اگر پاسخ ضربه دقیق فیلتر وینر برای این مسأله $h_w(n)$ نامیده شود، بر اساس قاعده تعامد سیگنال خطا ($e(n)$) بر داده ورودی ($x(n)$) :

$$E[e(n)x(n-j)] = r_{nn}(j) - \sum_{k=0}^{\infty} h_w(n)r_{xx}(j-k) = 0 \qquad (۵)$$

در رابطه (۵)، $r_{nn}(j)$ و $r_{xx}(j)$ به ترتیب تابع خودهبستگی نویز و سیگنال نویزی است. تابع خودهبستگی نویز از روی قاب‌های منطبق بر سکوت‌های مابین گفتار محاسبه می‌شود.

اگر بجای فیلتر وینر، فیلتر تقریبی و درجه دو $A_k(z)$ در

نظر گرفته شود، دیگر قاعده تعامد به طور دقیق برقرار نبوده و ضرب داخلی خطا و داده برابر مقدار مطلق صفر نخواهد بود. در این‌جا، می‌توان بر اساس مقدار ضرب داخلی داده و خطا، معیاری برای تعیین مرز باندها و تعیین ضرایب بهینه فیلتر $A_k(z)$ (یا معادلاً $a_{k,m}$ها) تعیین نمود. با فرض استقلال نویز و گفتار، از روی رابطه (٥)، کمیت:

$$E_k(j) = r_{nn}(j) - \sum_{m=0}^{2} a_{k,m}\, r_{xx}(j-m) + \sum_{m=0}^{2} a_{k,m}\, r_{nn}(j-m) \qquad (٦)$$

در نظر گرفته شده و حداقل شدن مقدار:

$$E_{ex}(i_k) = \sum_{j=0}^{2} E_k(j) \qquad (٧)$$

به عنوان معیار تعیین مرز هرقطعه از طیف قرار داده می‌شود. برای اطمینان از به حداقل رسیدن $E_{ex}(i_k)$، نقطه‌ای به عنوان مرز بالائی باند k–اُم یعنی i_k در نظر گرفته می‌شود که شرط:

$$\left| \frac{E_{ex}(i_k) - E_{ex}(i_k - M)}{E_{ex}(i_k - M)} \right| < \varepsilon \qquad (٨)$$

را برآورده نماید. ε یک عدد کوچک است که مقدار مناسب آن به صورت تجربی بدست می‌آید. M هم چنان تعیین می‌شود که حداقل پهنای ٣٠٠ هرتز برای هر فورمنت تضمین گردد.

لازم به ذکر است در روش ارائه شده برای قطعه‌بندی طیف سیگنال گفتار نویزی، باندهای مربوط به فورمنت اول، دوم، سوم و ... به ترتیب تعیین گردیده و پس از تعیین مرزها، فورمنت‌ها برابر با محل بیشینه طیف در هر باند در نظر گرفته می‌شود. بر اساس موارد فوق، الگوریتم تعیین محل فورمنت‌ها در جدول (١) خلاصه گردیده است.

جدول (١): الگوریتم تعیین محل فورمنت‌ها

مقداردهی اولیه:

$k = 1,\ i_{k-1} = 0,\ i_k = 1$

K = تعداد فورمنت‌های مورد نظر

مرحله اول:

برای باند k–اُم موارد زیر را انجام بده:

١) مقدار $r_{xx}(j)$ و $r_{nn}(j)$ را بر اساس فرمول (٤) محاسبه کن.

٢) بر اساس فرمول (٣) ضرائب فیلتر $A_k(z)$ را معین کن.

٣) با استفاده از فرمول (٦) و (٧) $E_{ex}(i_k)$ را بدست آور.

٤) اگر شرط (٨) برآورده می‌شود:

مرز بالائی باند k–اُم = i_k

محل ماکزیمم طیف در باند k = فورمنت k–اُم

به **مرحله دوم** برو.

$i_k = i_k + 1$(٥

مرحله دوم:

$k = k + 1$

$i_{k-1} = i_k$

اگر $k > K$ **پایان**.

در غیر این صورت به **مرحله اول** برو.

٣-٣- نحوه شکل دهی طیف سیگنال و تقویت اثر فورمنت‌ها

پس از مشخص شدن واکداری/بی‌واکی قاب و تعیین محل فورمنت‌ها (یا به بیان دقیق‌تر، تعیین باندهای مهم فرکانسی در سیگنال گفتار نویزی) به روش‌های گوناگون می‌توان نسبت به شکل‌دهی طیف و تضعیف اثر نویز اقدام نمود. در این تحقیق، روش‌های متعددی برای استفاده از اطلاعات بدست آمده در مورد واکداری/بی‌واکی و محل فورمنت‌ها مورد بررسی و پیاده‌سازی قرار گرفت [١]. آنچه در ادامه به عنوان روش فیلتر کردن و شکل‌دهی طیف سیگنال مطرح می‌شود، در واقع نتیجه ترکیب روش‌ها و بهینه‌سازی پارامترهای مختلف می‌باشد که بر اساس معیارهای کمی و کیفی، نتایج بهتری در کیفیت و قابلیت درک گفتار نویزی ایجاد نموده است.

در روش پیشنهادی این تحقیق، فرایند شکل‌دهی طیف سیگنال در دو مرحله صورت می‌پذیرد. ابتدا بسته به واکدار/بی‌واک بودن قاب مورد نظر و در صورت واکداری، با توجه به محل قله‌های طیف، فیلتر مناسب برای قاب مزبور تعیین و بر سیگنال اعمال می‌شود. فیلترهای مورد استفاده در این مرحله به شرحی که در بخش (٣-١) اشاره شد، بر اساس ویژگی‌های طیفی قاب‌های واکدار و بی‌واک طراحی گردیده است. در مورد واکدارها، فیلتر مورد استفاده فیلتری پایین‌گذر است که در محدوده زیر ١/٥ کیلوهرتز به شکل شانه‌ای[١٣] است به طوری که در محل هارمونیک‌های فرکانس گام دارای دامنه ١ و در فواصل بین آنها دارای ضریب افت ٠/٠١ است. در مورد بی‌واک‌ها نیز بر اساس ایده کلی تقویت باندهای فرکانسی پرانرژی و با بررسی طیف توان بی‌واک‌های پرتأثیر در قابلیت فهم گفتار (نظیر /s/ و /š/) [٤]، فیلتری بالاگذر با فرکانس قطع ٢٧٠٠ هرتز استفاده گردیده است.

مرحله دوم شکل‌دهی طیف نویز بر اساس اطلاعات بدست آمده در مورد محل فورمنت‌ها صورت می‌گیرد. بدین منظور با استفاده از مدل سیستم AR درجه دو برای هر فورمنت و ضرایبی که در حین اجرای الگوریتم قسمت (٣-٢-٢) برای هر قطعه از طیف گفتار حاصل می‌شود، تقریبی از طیف سیگنال گفتار تمیز محاسبه می‌گردد. در ادامه، یک فیلتر شکل‌دهنده

طیف (با پاسخ فرکانسی متناسب با تقریب حاصل از طیف سیگنال تمیز) روی سیگنال خروجی فیلترهای مرحله قبل اعمال می‌شود.

٤- پیاده‌سازی و ارزیابی کارآیی روش

در پیاده‌سازی و ارزیابی روش پیشنهادی، سیگنال‌های گفتار با فرکانس نمونه‌برداری Fs = 8 kHz مورد استفاده قرار گرفته و سیگنال با پنجره‌هایی از نوع همینگ و با طول ۲۰ میلی‌ثانیه (معادل ۱٦۰ نمونه) و همپوشانی ۵۰ درصد قاب‌بندی گردیده است. با توجه به فرکانس نمونه‌برداری ۸ کیلوهرتز (فرکانس قطع برابر با ٤ کیلوهرتز) و با فرض وجود یک فورمنت در هر یک کیلوهرتز پهنای باند گفتار [٤]، تعداد فورمنت‌ها برابر با $K = 4$ در نظر گرفته شد. برای محاسبه طیف هر قاب نیز از FFT-۱٦۰ نقطه‌ای استفاده گردید.

مطابق با آنچه در بخش (۳-۱) ارائه شد، اولین قدم در پردازش هر قاب، تعیین واکداری/بی‌واکی است. به خاطر اشکال‌های ناشی از نویز در پردازش حوزه فرکانس، تعیین واکداری/بی‌واکی با پردازش حوزه زمان صورت پذیرفته است؛ بدین منظور تابع خودهمبستگی قاب در اندیس‌های ۲٤ تا ۱۲۰ (محدوده منطقی برای دوره تناوب گام در فرکانس نمونه‌برداری ۸ کیلوهرتز) بررسی شده و به شرط وجود قله‌ای بزرگتر از $\delta = 0.7$، قاب مزبور واکدار و محل قله (I_0) به عنوان دوره تناوب گام در نظر گرفته می‌شود. در این حالت هارمونیک‌های فرکانس گام با فرمول $F_{0,k} = k \cdot Fs / I_0$ محاسبه و به عنوان محل قله‌های طیف سیگنال واکدار در طراحی فیلتر مورد استفاده در مرحله اول شکل‌دهی طیف مورد استفاده قرار می‌گیرد.

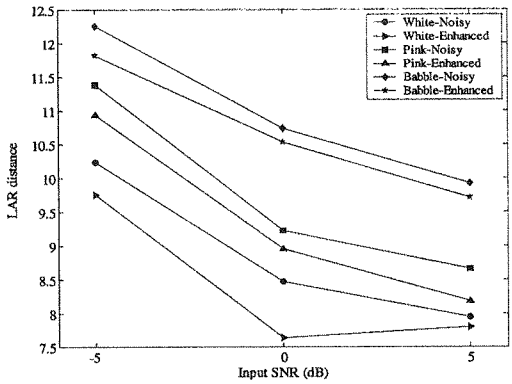
قدم بعدی در پردازش هر قاب، تعیین محل فورمنت‌ها بر اساس مطالب ارائه شده در بخش (۳-۲) است. در این راستا، الگوریتم ارائه شده در جدول (۱) پیاده‌سازی شده است. از جمله پارامترهای الگوریتم اخیر، M و ε می‌باشد. همان‌گونه که پیشتر هم بیان شد، M چنان انتخاب می‌شود که حداقل پهنای ۳۰۰ هرتز برای هر فورمنت تضمین گردد. با توجه به این که در FFT-۱٦۰ نقطه‌ای برای فرکانس نمونه‌برداری ۸ کیلوهرتز فاصله بین هر دو نمونه طیفی برابر با ۵۰ هرتز می‌باشد، مقدار مناسب برای M برابر با ٦ است. پارامتر ε نیز برای بررسی شرط همگرایی مورد استفاده قرار می‌گیرد و مقدار بهینه تجربی آن $\varepsilon = 0.001$ است.

٤-۱- ارزیابی کارآیی روش پیشنهادی

در این قسمت، عملکرد روش پیشنهادی توسط معیار کمی LAR-distance و آزمون کیفی MOS[14] مورد ارزیابی قرار گرفته است. همچنین از یک سری آزمون برتری[15] نیز برای بررسی بیشتر کارآیی بهره گرفته شد. در بررسی حاضر، دادگان سیگنال نویزی مورد استفاده مجموعه‌ای از سیگنال‌های نویزی شده (به صورت مصنوعی) و سیگنال‌های نویزی ضبط شده از محیط‌های واقعی بوده است. برای نویزی کردن سیگنال‌ها، ۵ جمله انتخاب شده از دادگان فارس‌دات در SNR = -5, 0, 5 dB با نویزهای سفید، صورتی و همهمه (از دادگان نویز Noisex [۱۵]) جمع شده است. محدوده SNR مورد بررسی با توجه به تعریف زمینه تحقیق، در مقادیر خیلی پایین در نظر گرفته شده است. علاوه بر ٤۵ جمله فوق، ۱۰ جمله ضبط شده در محیط‌های واقعی (نظیر داخل خودرو، فروشگاه، محیط دفتری با صدای پنکه و خیابان) نیز برای بررسی کارآیی سیستم مورد استفاده قرار گرفته است.

دلیل انتخاب LAR-distance به عنوان معیار کمی، همبستگی این معیار با نتایج آزمون‌های کیفی است [۱٦]. در سنجش میزان بهبود کیفیت گفتار با استفاده از معیار LAR-distance، فاصله میان سیگنال نویزی و سیگنال گفتار اولیه (تمیز) و نیز فاصله میان سیگنال خروجی روش و سیگنال گفتار اولیه مورد مقایسه قرار گرفته است. لازم به ذکر است به دلیل لزوم در اختیار بودن سیگنال تمیز اولیه، این سری از آزمون‌ها تنها بر روی سیگنال‌های نویزی‌شده (به طور مصنوعی) قابل اجراست.

متوسط نتایج حاصل از بررسی اخیر در شکل (۱) خلاصه گردیده است. همان‌گونه که پیداست، در مورد هر سه نوع نویز (سفید، صورتی و همهمه)، اعمال روش پیشنهادی به کاهش فاصله میان سیگنال خروجی با سیگنال گفتار اولیه و به بیان دیگر بهبود کیفیت سیگنال منجر گردیده است. بررسی دقیق‌تر این شکل حاکی از آن است که با حرکت از سمت نویز سفید به سوی نویزهای پایین‌گذر (و مشابه گفتار)، در SNR مساوی، اثر تخریبی نویز بر گفتار افزایش می‌یابد. همچنین، میزان بهبود



شکل (۱): مقدار LAR-Distance برای سیگنال‌های نویزی شده با سه نوع نویز در SNRهای مختلف و نیز سیگنال‌های خروجی سیستم

حاصل از اعمال روش پیشنهادی در SNRهای پایین به طور متوسط بیشتر است. در مجموع می‌توان گفت که عملکرد روش در مورد نویز سفید بهتر از دو نویز دیگر بوده است.

آزمون بعدی برای سنجش اثر اعمال روش پیشنهادی، آزمون MOS بوده است. برای انجام این آزمون از ۵ شنونده خواسته شد که به مجموعه‌ای از سیگنال‌های نویزی شده (در SNR = 0 dB) و نیز سیگنال‌های نویزی واقعی گوش فرا داده و بر اساس معیارهای آزمون MOS [٤] سیگنال‌ها را امتیازدهی نمایند.

نتایج حاصل از این بررسی در جدول شماره (۲) آمده است. اعداد مندرج در این جدول در واقع متوسط امتیازهایی است که ۵ شنونده مزبور به هر دسته ازسیگنال‌ها داده‌اند. نکته قابل ذکر آن که اگرچه بخاطر شدت نویزی بودن سیگنال‌ها، اعداد حاصل از آزمون MOS در امتیازهای پایین بوده، ولی در همه موارد اعمال روش پیشنهادی منجر به بالارفتن امتیاز MOS (به میزان ۰/۵ تا ۰/۷) گردیده است. اعداد این جدول حاکی از آن است که روش ارائه شده در بهبود قابلیت درک در مورد سیگنال‌های نویزی شده با نویز سفید به نحو مطلوب‌تر و در مورد سیگنال‌های نویزی واقعی (ضبط شده در محیط) به نحو کمرنگ‌تری عمل می‌نماید.

پس از بررسی کارآیی روش با دو ملاک و آزمون فوق، به عنوان آخرین روش مقایسه قابلیت درک سیگنال نویزی ورودی و سیگنال خروجی از آزمون برتری بهره گرفته شد. بدین منظور ۵ شنونده سیگنال‌های نویزی و خروجی سیستم را به صورت جفت‌جفت گوش داده و در هر مورد نسبت به برتری یکی از دو سیگنال از لحاظ قابلیت درک یا برابری آن دو اعلام نظر نموده‌اند. متوسط نتایج حاصل از اجرای این آزمون در جدول (۳) خلاصه گردیده است. نتایج این آزمون که به خوبی با آزمون‌های قبلی نیز همخوانی دارد، بر قابلیت روش ارائه شده در بهبود میزان درک گفتار تاکید می‌نماید.

جدول (۲): نتایج اجرای آزمون MOS برای سیگنال‌های نویزی و بهبودیافته (خروجی)

نوع سیگنال نویزی	MOS سیگنال نویزی	MOS سیگنال خروجی
نویزی شده با نویز سفید	۲/٤	۳/۱
نویزی شده با نویز صورتی	۲/۲	۲/۸
نویزی شده با نویز همهمه	۱/۸	۲/۳
نویزی ضبط شده در محیط	۱/۵	۲/۰

جدول (۳): نتایج اجرای آزمون برتری برای سیگنال‌های نویزی و بهبودیافته (خروجی)

نوع سیگنال نویزی	درصد برتری سیگنال نویزی	درصد برتری سیگنال خروجی	درصد برابری دو سیگنال
نویزی شده با نویز سفید	٪۳۱	٪٥٤	٪۱٥
نویزی شده با نویز صورتی	٪۳۳	٪٥٦	٪۱۱
نویزی شده با نویز همهمه	٪۳٤	٪٤۹	٪۱۷
نویزی ضبط شده در محیط	٪۲۹	٪٤۰	٪۳۱

در حین اجرای آزمون برتری، شنوندگان این نکته را یادآور گشته‌اند که اگرچه روش پیشنهادی در خیلی اوقات توانسته بخش قابل توجهی از نویز را حذف کند، لیکن در برخی از موارد سیگنال را دچار یک سری اشکال‌های جدید نموده و نمی‌توان گفت که سیگنال خروجی بهتر از سیگنال نویزی ورودی است. نمونه بارز این مسأله در مورد سیگنال‌های نویزی واقعی و به شکل بالا بودن درصد تساوی ورودی و خروجی منعکس گردیده است.

٤-۲- مقایسه کارآیی روش پیشنهادی با چند روش موجود

به منظور تکمیل ارزیابی‌های بخش (٤-۱) و سنجش کارآیی روش پیشنهادی از لحاظ میزان افزایش قابلیت درک از آزمون DRT[16] [٤] استفاده شد. در این مرحله، همچنین نتیجه عملکرد روش پیشنهادی با روش ارائه شده توسط Niederjohn [۷] (به عنوان روشی که به منظور خاص افزایش قابلیت درک سیگنال نویزی طراحی شده بود) و نیز با دو روش β-Order MMSE-STSA [٥] و OM-LSA [٦] (به عنوان دو روش جدید بهسازی طیفی سیگنال گفتار) مورد مقایسه قرار گرفت.

مجموعه کلمات فارسی تهیه شده برای آزمون DRT فارسی [۳] در محیط اتاق قرائت و ضبط و در ادامه، به عنوان ورودی روش‌های مورد مقایسه استفاده شد. برای ایجاد شرائط نویزی شدید که در آن قابلیت درک گفتار تا حد زیادی با مشکل روبرو باشد، ضبط سیگنال در وضعیت ترکیب شده از صدای شدید کولر و همهمه جمعیت صورت پذیرفت.

آزمون DRT با حضور ۵ شنونده برگزار و متوسط درصد تشخیص درست کلمات در جدول شماره (٤) خلاصه شده است. همان‌گونه که انتظار می‌رود سیگنال ضبط شده در محیط با نویز شدید از قابلیت درک خوبی برخوردار نیست. هم روش پیشنهادی و هم سه روش مورد مقایسه با آن توانسته‌اند (تا حدی) میزان قابلیت درک گفتار را افزایش دهند. دیده می‌شود

که روش‌های β-Order MMSE-STSA و OM-LSA با آن که به منظور خاص افزایش قابلیت درک گفتار طراحی نشده‌اند، ولی در این زمینه نیز عملکرد خوبی دارند. روش پیشنهادی این تحقیق، عملکرد برتری را نسبت به کلیه روش‌ها ارائه می‌دهد.

جدول (۴): نتایج اجرای آزمون DRT بر روی کلمات نویزی و خروجی روش‌های Niederjohn، β-Order MMSE-STSA، OM-LSA و روش پیشنهادی

	Noisy Input	Nieder-john	β-MMSE	OM-LSA	روش پیشنهادی
%DRT	٪۵۳	٪۷۲	٪۶۷	٪۷۴	٪۸۱

۵- جمع بندی

این مقاله با معرفی مسأله بهسازی گفتار و مطالعه عوامل مؤثر در قابلیت درک آغاز شد. ملاحظه شد که نویز، انعکاس، کم بودن پهنای باند و نیز ساختار واجی گفتار از پارامترهای مؤثر در قابلیت درک گفتار می‌باشد. در این میان نویز به عنوان اصلی‌ترین عامل تخریب سیگنال گفتار حائز اهمیت خاص است. در این تحقیق، به نویز به دید اصلی‌ترین عامل تخریب سیگنال و کاهش قابلیت درک نگریسته شده و راه‌حلی برای بهبود قابلیت درک گفتار نویزی ارائه شد.

در روش ارائه شده پس از تعیین واکداری/بی‌واکی هر قاب گفتار، فیلتر مناسب برای قاب مزبور معین و بر روی سیگنال نویزی اعمال می‌گردد. روش پیشنهادی همچنین دارای یک الگوریتم تعیین محل فورمنت می‌باشد. با تعیین محل فورمنت‌ها و با استفاده از مدل تشدیدکننده درجه دوم برای هر فورمنت فیلتری متناسب با طیف سیگنال طراحی و بر روی سیگنال خروجی مرحله قبل اعمال می‌شود. در ادامه، خروجی روش ارائه شده توسط معیار کمی LAR-distance و آزمون‌های MOS، برتری و DRT با سیگنال نویزی ورودی و نیز خروجی‌های سه روش معروف و متداول بهسازی و بهبود قابلیت درک گفتار مقایسه شده و برتری سیگنال خروجی روش پیشنهادی مورد تأیید قرار گرفت.

مراجع

[۱] ابوطالبی، حمیدرضا؛ **بهبود قابلیت فهم گفتار نویزی با استفاده از روش‌های بهسازی گفتار**، گزارش طرح پژوهشی، دانشگاه یزد، ۱۳۸٤.

[۲] منسوب بصیری، مهدی؛ احدی، سید محمد؛ **"کاهش نویز سیگنال گفتار و افزایش قابلیت فهم کلام با استفاده از شکل‌دهی طیف نویز"**، مجموعه مقالات دهمین کنفرانس سالانه انجمن کامپیوتر ایران، بهمن ۱۳۸۳.

[۳] ابوطالبی، حمیدرضا؛ فغانی، فرهاد؛ **"ارائه دادگان فارسی برای آزمون DRT جهت سنجش قابلیت فهم سیگنال گفتار"**، مجموعه مقالات چهاردهمین کنفرانس مهندسی برق ایران، اردیبهشت ۱۳۸۵.

[٤] Deller, J. R.; Hansen, J. H. L.; Proakis, J. G.; Discrete-Time Processing of Speech Signals, 2nd edition, IEEE Press, 2000.

[٥] You, C. H.; Koh, S. N.; Rahardja, S.; "β-Order MMSE Spectral Amplitude Estimation for Speech Enhancement", IEEE Signal Processing Letters, Apr. 2002.

[٦] Cohen, I.; "On Speech Enhancement under Signal Presence Uncertainty Using Spectral Amplitude Estimator", IEEE Signal Processing Letters, Apr. 2002.

[٧] Niederjohn, R. J.; Grotelueschen, J. H.; "The Enhancement of Speech Intelligibility in High Noise Levels by High-Pass Filtering Followed by Rapid Amplitude Compression", IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 24, no. 4, pp. 277-282, Aug. 1976.

[٨] Niederjohn, R. J.; Svoren, T. J.; Heinen, J. A.; "Intelligibility Enhancement of Noise Corrupted Speech Based on Formant Tracking Involving Pre-filtering", Proc. ICASSP, pp. 1336-1339, 1992.

[٩] Kitamura, T; et al.; "Designing Modulation Filters for Improving Intelligibility in Reverberant Environments", Proc. ICSLP, vol. 3, pp. 386-389, 2000.

[۱۰] Rodman, J.; "The BRAIN Model of Intelligibility in Business Telephony", Polycom Tech. Report, Jan. 2003.

[۱۱] Haykin, S.; Adaptive Filter Theory, 3rd edition, Prentice-Hall, 1996.

[۱۲] Rodman, J.; "The Effect of Bandwidth on Speech Intelligibility", Polycom Tech. Report, Jan. 2003.

[۱۳] Klatt, D. H.; "Review of Text to Speech Conversion for English", Journal of Acoustical Society of America, vol. 82, pp. 737-793, 1987.

[۱٤] Welling, L.; Ney, H.; "Formant Estimation for Speech Recognition", IEEE Trans. on Speech and Audio Processing, vol. 6, pp. 36-48, 1998.

[۱٥] http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Data/ noisex.html

[۱٦] Quackenbush, S. R.; Barnwell III, T. P.; Clements, M. A.; Objective Measures of Speech Quality, Prentice-Hall, Englewood Cliffs, NJ, 1988.

زیرنویس‌ها

[1] Speech Enhancement
[2] Intelligibility
[3] Pre-processing
[4] Post-processing
[5] Minimum Mean Square Error-Short Time Spectral Amplitude
[6] Optimally Modified-Log Spectral Amplitude
[7] Frame
[8] Formant
[9] Reverberation/Echo
[10] Phonemic Structure
[11] Linear Prediction Coding
[12] Notch Filter
[13] Comb
[14] Mean Opinion Score
[15] Preference
[16] Diagnostic Rhyme Test