Computational Intelligence in Electrical Engineering
Vol. 11, No. 1, 2020
Research Paper

Single-channel Speech Enhancement using the Combination of Exponential Deterministic Model and t Location-scale Stochastic Model

Zahra Amini[1], Neda Faraji[2]

[1] MSc Student, Dept. of Electrical Engineering, Imam Khomeini International University, Qazvin, Iran
nfaraji@aut.ac.ir
[2] Assistant Professor, Dept. of Electrical Engineering, Imam Khomeini International University, Qazvin, Iran
zahra.amini56@yahoo.com

Abstract:
Most speech enhancement algorithms focus on obtaining an estimator relying on stochastic models. In this paper, a minimum mean-square error (MMSE) estimator under a stochastic–deterministic model is proposed where a heavy-tail distribution called t-Location-Scale (tls) is used for modeling Discrete Fourier Transform coefficients of clean speech signals and exponential and sinusoidal models are employed as deterministic models. In the exponential model, the frequency and damping coefficient are estimated by using the Matrix Pencil method. Also, in previous studies, the number of exponential components in the deterministic model for stochastic-deterministic speech enhancement algorithm has been considered to be one. In this paper, the corresponding exponential model is developed to have an arbitrary number of exponential components. The speech enhancement experiments are performed in three modes, exponential-Gaussian (the first proposed method), exponential-tls (the second proposed method), and sinusoidal-Gaussian. Comparisons are made with the exponential-Gaussian method (with only one exponential component), as well as with the Weiner and tls stochastic estimators. The implementation results in the presence of six noise types from Noisex-92 dataset show that the two proposed methods improve the segSNR values and have quite similar PESQ values comparing with the stochastic based speech enhancement methods.

Keywords: Speech Enhancement, t Location-scale Probability Density Function, Wiener Filter, Minimum Mean Square Error, Exponential Deterministic Model, Sinusoidal Model.

http://dx.doi.org/10.22108/isee.2019.114459.1171

مقاله پژوهشی

بهسازی گفتار تک‌کاناله با استفاده از ترکیب مدل قطعی نمایی و مدل تصادفی *t* Location-Scale

زهرا امینی[۱]، ندا فرجی[۲]

۱- دانشجوی کارشناسی ارشد، گروه مهندسی برق - دانشگاه بین‌المللی امام خمینی (ره) - قزوین - ایران
zahra.amini56@yahoo.com
۲- استادیار، گروه مهندسی برق - دانشگاه بین‌المللی امام خمینی (ره) - قزوین - ایران
nfaraji@eng.ikiu.ac.ir

چکیده: بیشتر روش‌های بهسازی گفتار، تخمینگری کاملاً متکی به مدل تصادفی گفتار ارائه می‌دهند. در این مقاله، یک تخمینگر کمترین میانگین مربعات خطا تحت یک مدل قطعی - تصادفی پیشنهاد می‌شود که در آن از یک توزیع دنباله - سنگین به نام *t* location-scale (*t*ls) برای مدل‌کردن ضرایب تبدیل فوریه گسسته گفتار تمیز و از مدل نمایی و سینوسی به‌عنوان مدل قطعی استفاده شده است. در مدل نمایی به‌کاررفته، تخمین فرکانس و ضریب میرایی به روش ماتریس پنسِل انجام می‌شود. همچنین، در پژوهش‌های قبلی تعداد مؤلفه‌های نمایی در ساخت مدل قطعی برای بهسازی گفتار، یک در نظر گرفته شده است که در این مقاله، مدل نمایی به تعداد دلخواه مؤلفه‌های نمایی بسط داده می‌شود. پیاده‌سازی‌ها در سه حالت ترکیبی نمایی - گاوسی (روش پیشنهادی نخست)، نمایی - *t*ls (روش پیشنهادی دوم) و سینوسی - گاوسی انجام شده‌اند و با روش موجود نمایی – گاوسی (تنها با یک مؤلفهٔ نمایی) و تخمینگرهای تصادفی وینر و مبتنی بر *t*ls مقایسه می‌شوند. نتایج پیاده‌سازی در حضور شش نویز از مجموعه دادهٔ نویز noisex-92 نشان می‌دهند که دو روش پیشنهادی در قیاس با روش‌های مبتنی بر مدل تصادفی صرف، به بهبود معیار نسبت سیگنال به نویز قطعه‌ای منجر شده‌اند و در ارزیابی ادراکی کیفیت گفتار عملکرد نسبتاً برابری دارند.

واژه‌های کلیدی: بهسازی گفتار، تابع چگالی احتمال *t* Location-Scale، فیلتر وینر، کمترین میانگین مربعات خطا، مدل قطعی نمایی، مدل سینوسی

۱- مقدمه

با رشد روزافزون استفاده از سیستم‌های گفتاری در کاربردهای علمی و روزمره، نیاز به حفظ کیفیت گفتار، امری اجتناب‌ناپذیر مطرح شده است. شرایط ایده‌آل و عاری از نویزی که در کارها و شبیه‌سازی‌های آزمایشگاهی در نظر گرفته می‌شوند، در بسیاری از کاربردهای واقعی به‌طور جدی نقض شده‌اند و برقراری آنها زیر سؤال می‌رود؛ ازاین‌رو، مبحث بهسازی گفتار، یکی از ضرورت‌های کاربردی و عملی، از زمینه های فعال تحقیقاتی در سال های اخیر بوده است.

بهسازی گفتار، فرایند بازیابی گفتار تمیز از سیگنال نویزی گفتار است و در مواردی به کار می‌رود که سیگنال گفتار از نویز، انعکاس یا سایر عوامل مخرب تأثیر گرفته است.

در سه دهه اخیر، استراتژی‌های مختلفی برای بهسازی گفتار در حضور نویز جمع‌شونده پیشنهاد شده‌اند [۱]. بیشتر

[۱] تاریخ ارسال مقاله: ۱۳۹۷/۰۹/۲۰
تاریخ پذیرش مقاله: ۱۳۹۸/۰۸/۱۱
نام نویسندهٔ مسئول: ندا فرجی
نشانی نویسندهٔ مسئول: ایران، قزوین، دانشگاه بین‌المللی امام خمینی (ره)، دانشکده فنی، گروه مهندسی برق

این استراتژی‌ها از الگوریتم‌های بهسازی گفتار در حوزهٔ فرکانس استفاده می‌کنند که باوجود پیچیدگی محاسباتی کمتر، نتایج بهبود کیفیت چشمگیری را به همراه دارند. روش متداول تفریق طیفی [۲] که محاسبات ریاضی ساده دارد، فیلتر وینر و تغییرات آن مانند فیلترینگ تکراری وینر [۳] از روش‌های بهسازی در حوزهٔ فرکانس‌اند.

گروه دیگر از روش های بهسازی در حوزهٔ فرکانس روش‌های مبتنی بر مدل آماری‌اند. در این روش‌ها عموماً از روش بیزین برای تخمین سیگنال بهبودیافته استفاده می‌شود. ازجمله روش‌های آماری متداول می‌توان به تخمینگر کمترین میانگین مربعات خطا[۱] (MMSE) [۲]، تخمینگر مبتنی بر لگاریتم کمترین میانگین مربعات خطا (Log-MMSE) [٤]، بیشینهٔ درست‌نمایی[۲] (ML) [٥] و تخمینگر بیشینهٔ احتمال پسین[۳] (MAP) [٦] اشاره کرد.

تخمینگر کمترین میانگین مربعات خطا و بیشینهٔ درست نمایی عمدتاً در حوزهٔ تبدیل فوریه گسسته استفاده می‌شوند و با استفاده از این روش‌ها دامنه یا ضرایب مختلط تبدیل فوریه سیگنال گفتار از ضرایب مختلط تبدیل فوریه گسسته گفتار نویزی تخمین زده می‌شود.

در تخمینگرهای کمترین میانگین مربعات خطا، یک تابع هزینهٔ استاندارد ریاضی یا یک معیار پذیرفتنی بهینه می‌شود تا یک تابع بهره غیرخطی برای اصلاح ضرایب تبدیل فوریه گسسته سیگنال نویزی به دست آید. برای پیداکردن تخمینگرهای کمترین میانگین مربعات خطا و بیشینهٔ درست نمایی به دو تابع چگالی احتمال نیاز است؛ تابع چگالی احتمال پیشین (تابع چگالی احتمال سیگنال تمیز) و تابع چگالی احتمال سیگنال نویز.

با این فرض که ضرایب تبدیل فوریه گسسته سیگنال گفتار تمیز و نویز هر دو گاوسی اند، تخمینگر کمترین میانگین مربعات خطا به دست آمده است [۲]. در پژوهش های انجام‌شده دربارهٔ تابع چگالی احتمال سیگنال گفتار تمیز در حوزهٔ فرکانس و زمان، نشان داده شد که قسمت‌های حقیقی و موهومی ضرایب تبدیل فوریه گسسته سیگنال تمیز یک توزیع سوپرگاوسی[٤] دارند (یک قلهٔ واضح‌تر و دنباله‌های سنگین در مقایسه با گاوسی) [٦]. تا به امروز عمدتاً فرض بر این بوده است که نویز دارای تابع چگالی احتمال گاوسی و سیگنال گفتار تمیز دارای تابع چگالی احتمال لاپلاس [۷]، گاما [۸]، گامای تعمیم‌یافته[٥] [۹] و [t location-scale ۱۰] است. یک اصلاح دیگر برای بهبود الگوریتم های بهسازی گفتار مبتنی بر روش های آماری، در نظر گرفتن احتمال حضور[٦] (غیاب) گفتار است [۱].

استفاده از مدل قطعی در روش‌های بهسازی گفتار نخستین‌بار در سال ۱۹۸۰ مطرح شد [۱۱]. در این مقاله، سیگنال گفتار تمیز با یک تابع نمایی با دامنه و فاز نامشخص نمایش داده شد. پس از آن در سال ۱۹۹۳، در چارچوب فیلتر وینر، یک الگوریتم بهسازی گفتار ارائه شد که هم‌زمان هر دو جزء گفتار قطعی و تصادفی را در نظر می گیرد [۱۲]. در این مقاله، سیگنال گفتار به دو بخش مصوت (بخش پریودیک سیگنال گفتار) و نامصوت (بخش غیرپریودیک سیگنال گفتار) تقسیم شد و قسمت مصوت گفتار با مدل قطعی و قسمت نامصوت گفتار نیز تصادفی در نظر گرفته شد. در سال ۲۰۰۷، هندریکس و همکارانش برای ضرایب تبدیل فوریه گفتار، مدل تصادفی یا قطعی در نظر گرفتند و سپس تخمین کمترین میانگین مربعات خطا را مبتنی بر هر دو تصمیم نرم و سخت در انتخاب بین دو مدل تصادفی و قطعی گفتار به دست آوردند [۱۳]؛ البته در مدل نویز به‌علاوهٔ هارمونیک از سیگنال گفتار، ضرایب تبدیل فوریه سیگنال تمیز در هر لحظه از هر دو مدل تصادفی و قطعی پیروی می‌کنند که با یک توزیع با میانگین غیرصفر نمایش داده می‌شوند [١٤]. مک‌کالوم و همکاران، این ایده را در سال ۲۰۱۲ استفاده کردند. در این مقاله بر خلاف کارهای پیشین که برای نویز یک مدل تصادفی در نظر می گرفتند، پژوهشگران یک مدل نویز تصادفی – قطعی در نظر گرفتند [١٥]. در سال ۲۰۱۳ نیز مک کالوم و همکارانش یک الگوریتم بیزین تحت مدل گفتار تصادفی – قطعی با اطلاعات پیشین (استفاده از اطلاعات فریم قبلی برای تخمین فرکانس هر فریم) را پیشنهاد کردند. در این الگوریتم، میانگین توزیع غیرصفر در نظر گرفته می‌شود [١٦].

به‌تازگی نیز روش بهسازی مبتنی بر شبکهٔ عصبی عمیق[۷] ارائه شده است که از داده های آموزشی برای یافتن نگاشت بین داده‌های گفتار نویزی و تمیز استفاده می‌کند [۱۷].

در این مقاله، روش بهسازی گفتار مبتنی بر مدل قطعی – تصادفی تعمیم داده می شود که در [۱۳] بیان شده است. در [۱۳] از توزیع تصادفی گاوسی و لاپلاس به‌عنوان مدل تصادفی و از مدل نمایی به‌عنوان مدل قطعی استفاده کردند. در این مقاله از توزیع جدیدی به نام t location-scale به‌عنوان مدل تصادفی استفاده می شود. همچنین، مدل نمایی استفاده‌شده در [۱۳]، با افزایش پارامتر P و تغییر روش تخمین فرکانس تعمیم داده می‌شود.

ساختار مقالهٔ حاضر به شرح زیر است: در بخش دوم، مقولهٔ مدل‌سازی سیگنال گفتار با استفاده از مدل قطعی – نمایی به همراه روش‌های مختلف تخمین فرکانس و مقایسهٔ آنها بررسی می شود. در بخش سوم، چند روش بهسازی گفتار تحت مدل تصادفی و مدل تصادفی – قطعی در حضور نویز جمع‌شونده معرفی می‌شوند. در بخش چهارم مقاله، روش پیشنهادی مطرح شده است. در بخش پنجم، روش های بهسازی گفتار پیشنهادشده با روش های پیشین در این حوزه مقایسه می شوند و در بخش پایانی، نتیجه‌گیری کلی از مقاله ارائه می‌شود.

۲- مدل‌سازی سیگنال گفتار با استفاده از مدل قطعی نمایی

مدل قطعی نمایی یک سیگنال، شامل جمع P تابع تک فرکانس مختلط مطابق رابطهٔ (۱) است:

$$y[n]=\sum_{i=1}^{P}\frac{A_i}{2}e^{j\phi_i}e^{(d_i+j2\pi f_i)n}, \qquad (۱)$$

در این رابطه، n اندیس نمونه های زمانی است و f_i, d_i, ϕ_i, A_i نیز به ترتیب دامنه، فاز، ضریب میرایی و فرکانس iامین تابع فرکانس مختلط را نشان می دهند. روش های مختلفی برای تخمین پارامترهای فرکانس و ضریب میرایی وجود دارند. روش پرونی[۸] از ساده ترین روش‌هاست که با افزایش تعداد فرکانس های مدل، دقت پایین‌تری خواهد داشت [۱۸]. روش‌های مک‌برایـد[۹]، ماتریس پِنسِل[۱۰] و حداقل مربعات کل[۱۱] در مراجع [۱۹]، [۲۰] و [۲۱]، روش‌های مؤثر در تخمین فرکانس و ضریب میرایی معرفی شده‌اند. در این بخش، این سه روش در تعداد نمونه های مختلف از سیگنال، تعداد فرکانس های تخمینی و نسبت سیگنال به نویزهای مختلف (از صفر تا ۴۰ دسی بل) با یکدیگر مقایسه می شوند. همچنین، برای مقایسهٔ روش‌های مختلف تخمین فرکانس و ضریب میرایی از معیار میانگین قدر مطلق خطا به‌ترتیب مطابق با روابط (۲) و (۳) استفاده می‌شود.

$$MAE_f=\frac{1}{P}\sum_{p=1}^{P}\left|\hat{f}_p-f_p\right|, \qquad (۲)$$

$$MAE_d=\frac{1}{P}\sum_{p=1}^{P}\left|\hat{d}_p-d_p\right|, \qquad (۳)$$

در این دو رابطه، P تعداد توابع نمایی در معادلهٔ (۱) و $d_p, f_p, \hat{d}_p, \hat{f}_p$ نیز به ترتیب فرکانس تخمینی، ضریب میرایی تخمینی، فرکانس واقعی و ضریب میرایی واقعی pامین مؤلفهٔ فرکانس مختلط را نشان می‌دهند.

سیگنال ورودی مطابق با رابطهٔ (۱) با فرکانس، ضریب میرایی، دامنه، فاز و تعداد نمونه های مختلف ساخته شده است و به هرکدام از روش‌های تخمین داده می‌شود. سپس مطابق با روابط (۲) و (۳) مقدار میانگین قدر مطلق خطای تخمین محاسبه می‌شود.

شکل (۱) اثر تعداد فرکانس بر تخمین فرکانس و ضریب میرایی را وقتی تعداد نمونه‌ها برابر ۵۰۰ است، در سه روش مک برایـد (StMcB)، ماتریس پِنسِل (MatPen) و حداقل مربعات کل (Total LS) نشان می‌دهد. در شکل (۱.الف) و (۱.ج)، به‌ترتیب خطای تخمین فرکانس و خطای تخمین ضریب میرایی، وقتی تعداد فرکانس‌ها ۱۰ و در شکل (۱.ب) و (۱.د)، خطای تخمین فرکانس و خطای تخمین ضریب میرایی، وقتی تعداد فرکانس‌ها ۲۰ باشد، نشان داده شده است. طبق این شکل با افزایش تعداد فرکانس، خطای تخمین در روش ماتریس پِنسِل کمتر از بقیه روش‌های تخمین است.

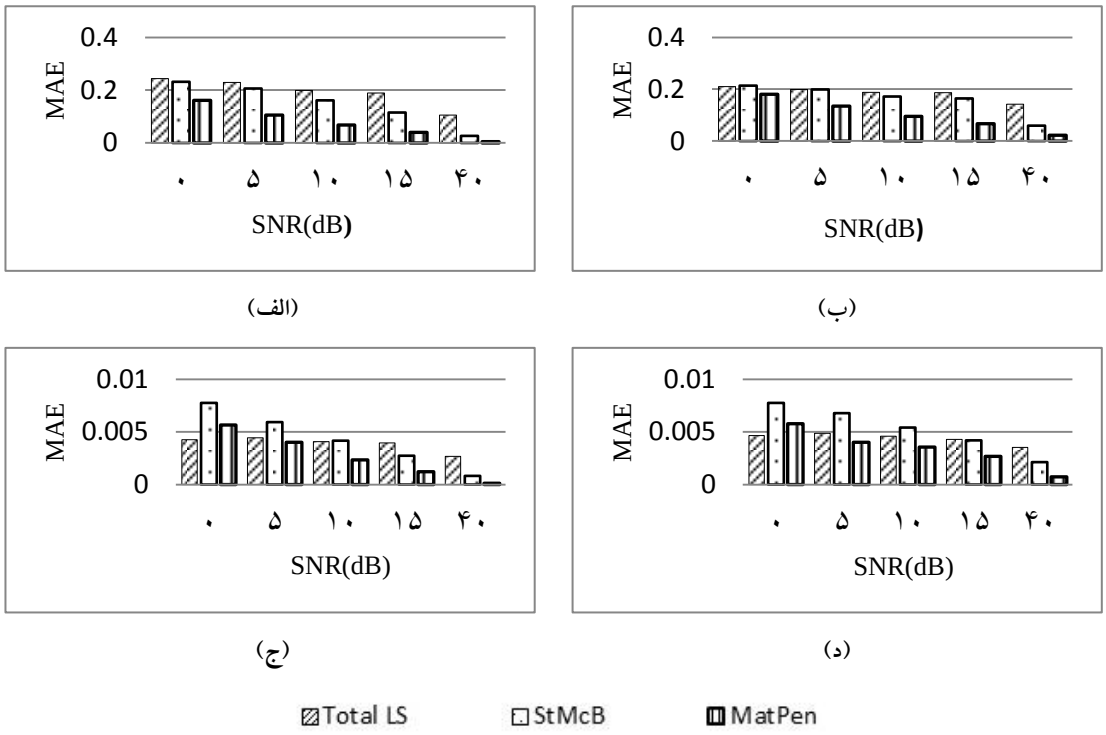
شکل (۲) اثر تعداد نمونه ها بـر تخمـین فرکـانس و ضریب میرایی را وقتی تعداد فرکانس ها برابر ۲۰ اسـت، در سه روش مک‌براید، ماتریس پِنسِل و حداقل مربعـات کل نشان می‌دهد. در شکل (۲.الف) و (۲.ج)، بـه‌ترتیـب خطای تخمین فرکانس و خطای تخمین ضریب میرایـی، وقتی تعداد نمونـه هـا ۵۰۰ و در شـکل (۲.ب) و (۲.د)، خطای تخمین فرکانس و خطای تخمین ضریب میرایـی، وقتی تعداد نمونه ها ۱۰۰۰ باشد، نشان داده شـده اسـت. مطابق این شکل، روش ماتریس پِنسِل با افـزایش تعـداد نمونه ها، به خطای تخمین کمتـری در مقایسـه بـا سـایر روش ها منجر می شود. طبق مشاهدات مذکور که حـاکی از قدرت روش ماتریس پنسل است، در بخش های بعدی برای تخمـین فرکـانس و ضـریب میرایـی از ایـن روش استفاده می‌شود.

۳- بهسازی گفتار در نویز جمع شونده مبتنـی بر روش‌های تصادفی و قطعی - تصادفی

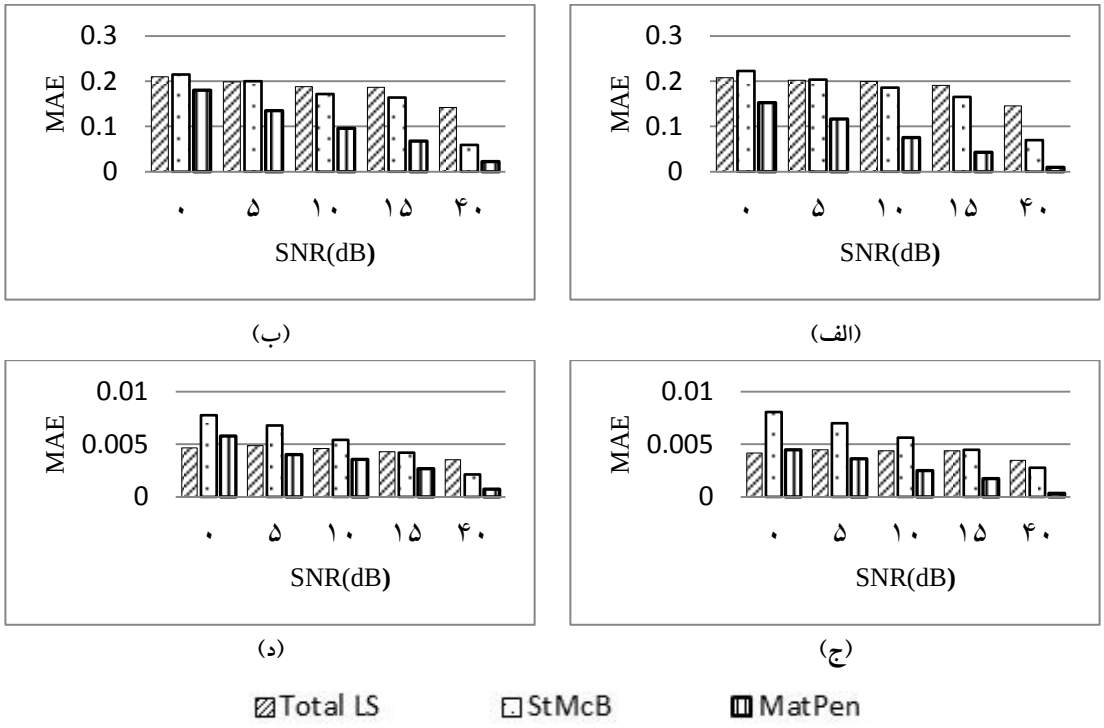
اگر $s(n)$ سـیگنال تمیـز و $w(n)$ نـویز جمـع شـونده باشد، سیگنال نویزی $y(n)$ طبق رابطهٔ (٤) به دست می‌آید:

$$y(n)=s(n)+w(n). \qquad (٤)$$

با فرض ناهمبسته‌بودن نویز با سیگنال اصلی، در حـوزهٔ فوریه رابطهٔ (۵) برقرار است.



شکل (۱): بررسی اثر تعداد فرکانس‌ها بر خطای تخمین فرکانس و ضریب میرایی در روش‌های ماتریس پِنسِل (MatPen ، مک‌براید (StMcB) و حداقل مربعات کل (Total LS) در ۵۰۰ نمونه (الف): خطای تخمین فرکانس با ۲۰ فرکانس، (ب): خطای تخمین فرکانس با ۱۰ فرکانس، (ج): خطای تخمین ضریب میرایی با ۲۰ فرکانس، (د): خطای تخمین ضریب میرایی با ۱۰ فرکانس.



شکل (۲): بررسی اثر تعداد نمونه‌ها بر خطای تخمین فرکانس و ضریب میرایی در روش‌های ماتریس پنسِل (MatPen)، مک‌براید (StMcB) و حداقل مربعات کل (Total LS) در ۲۰ فرکانس (الف): خطای تخمین فرکانس با ۱۰۰۰ نمونه، (ب): خطای تخمین فرکانس با ۵۰۰ نمونه، (ج): خطای تخمین ضریب میرایی با ۱۰۰۰ نمونه، (د): خطای تخمین ضریب میرایی با ۵۰۰ نمونه.

$$Y(k,i)=S(k,i)+W(k,i), \qquad (۵)$$

$Y(k,i)$، $S(k,i)$ و $W(k,i)$ به‌ترتیب تبدیل فوریه سیگنال نویزی، سیگنال تمیز و نویزند. k و i نیز به‌ترتیب نشان‌دهندهٔ اندیس فرکانس و شماره فریم‌اند.

۳-۱- بهسازی گفتار با فرض مدل تصادفی برای گفتار

در این بخش، بهسازی گفتار با استفاده از معیار کمینهٔ میانگین مربعات خطا و با فرض تابع چگالی احتمال گاوسی و location-scale t برای قسمت حقیقی و موهومی ضرایب فوریه زمان کوتاه سیگنال گفتار تمیز معرفی می‌شود. همچنین، فرض می‌شود قسمت حقیقی و موهومی ضرایب فوریه زمان کوتاه از هم مستقل‌اند و ضرایب فوریه زمان کوتاه برای سیگنال نویز نیز از توزیع گاوسی مختلط با میانگین صفر تبعیت می‌کند.

۳-۱-۱- بهسازی با فرض مدل تصادفی گاوسی

تحت مدل تصادفی و استفاده از این فرض که ضرایب تبدیل فوریه گسسته سیگنال تمیز، توزیع مختلط گاوسی با میانگین صفر دارند، ضرایب تبدیل فوریه گسسته سیگنال نویزی، توزیع مختلط گاوسی با میانگین صفر خواهند داشت. سیگنال گفتار تمیز با رابطهٔ (۶) تخمین زده می شود [۲]:

$$\hat{s}(k,i)=E\{S(k,i)\,|\,y(k,i)\}=\frac{\xi(k,i)}{1+\xi(k,i)}y(k,i), \qquad (۶)$$

$\xi(k,i)$ مقدار سیگنال به نویز پیشین است که در بخش بعد روش تخمین آن توضیح داده می‌شود.

۳-۱-۲- بهسازی با فرض مدل تصادفی t location-scale

در این مدل فرض می شود ضرایب حقیقی و موهومی تبدیل فوریه زمان کوتاه سیگنال تمیز دارای توزیع t

location-scale و ضرایب تبدیل فوریه زمان کوتاه سیگنال نویز دارای توزیع مختلط گاوسی با میانگین صفرند. تابع چگالی احتمال t location-scale برای متغیر تصادفی S با ν درجه آزادی و میانگین صفر به صورت زیر تعریف می‌شود [۱۰]:

$$f_S(s)=\frac{\Gamma\left((\nu+1)/2\right)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1+s^2/(b^2\nu)\right)^{\frac{-(\nu+1)}{2}}, \qquad (۷)$$

$\nu>0$، b نشان دهندهٔ مقیاس، $f(.)$ نشان دهندهٔ تابع چگالی احتمال و $\Gamma(.)$ تابع گاما است.

تحت مدل گفتار t location-scale، ضرایب مختلط تبدیل فوریه زمان کوتاه سیگنال گفتار تمیز، با ترکیب تخمین های قسمت های حقیقی و موهومی به فرم رابطهٔ (۸) تخمین زده می‌شوند [۱۰]:

$$\begin{aligned}\hat{s}(k,i)&=E\{S(k,i)\,|\,y(k,i)\}\\&=E\{S_R\,|\,y_R\}+jE\{S_I\,|\,y_I\},\end{aligned} \qquad (۸)$$

در این رابطه، اندیس‌های R و I به‌ترتیب نشان‌دهندهٔ قسمت حقیقی و موهومی‌اند و $j=\sqrt{-1}$ است. قسمت حقیقی طبق رابطهٔ (۹) تخمین زده می شود و جزء موهومی نیز طبق رابطهٔ مشابه و با جایگزینی اندیس R با I به دست می‌آید[۱۰].

$$E\{S_R\,|\,y_R\}=\frac{\sum_{j=0}^{\infty}\frac{1}{j!}\left(\frac{\varsigma_R}{2}\right)^j\Gamma\left(\frac{\nu+1}{2}\right)}{\sum_{j=0}^{\infty}\frac{1}{j!}\left(\frac{\varsigma_R}{2}\right)^j\Gamma\left(\frac{\nu+1}{2}\right)}\times \frac{\psi\left(\frac{\nu+1}{2},\frac{\nu}{2}-j;\frac{\xi_R(\nu-2)}{2}\right)}{\psi\left(\frac{\nu+1}{2},\frac{\nu}{2}+1-j;\frac{\xi_R(\nu-2)}{2}\right)}y_R, \qquad (۹)$$

در این رابطه $\psi(.)$ تابع Tricomi's hypergeometric است و $\xi(k,i)$ و $\varsigma(k,i)$ به‌ترتیب نشان‌دهندهٔ مقدار سیگنال به نویز پیشین و پسین در بین فرکانسی[۱۲] k و فریم iام هستند که در ادامه تعریف می‌شوند.

$$\xi(k,i)=\frac{E\left[S(k,i)^2\right]}{E\left[W(k,i)^2\right]}=\frac{\sigma_S^2(k,i)}{\sigma_W^2(k,i)}, \qquad (۱۰)$$

در این رابطه، σ_S^2 و σ_W^2 به‌ترتیب واریانس گفتار تمیز و نویز را نشان می دهند. هرچند واریانس سیگنال گفتار تمیز در دسترس نیست، افرایم و همکارانش روش تصمیم‌گیری مستقیم (Decision-Directed) را برای محاسبهٔ آن پیشنهاد دادند که مطابق با رابطهٔ (۱۱) است [۱]:

$$\xi(k,i)=\alpha\frac{\left|\hat{s}(k,i-1)\right|^2}{\sigma_W^2(k,i)}+(1-\alpha)\times\max\left(\varsigma(k,i)-1,0\right), \qquad (۱۱)$$

α ثابت هموارسازی است و $|.|$ عملگر قدر مطلق را نشان می دهد. $\varsigma(k,i)$، یعنی مقدار سیگنال به نویز پسین در بین فرکانسی kام و فریم iام به‌راحتی با رابطهٔ (۱۲) محاسبه می‌شود:

$$\varsigma(k,i)=\frac{E\left[Y(k,i)^2\right]}{E\left[W(k,i)^2\right]}=\frac{\sigma_Y^2(k,i)}{\sigma_W^2(k,i)}. \qquad (۱۲)$$

۳-۲- بهسازی گفتار به روش قطعی - تصادفی

در این بخش، یک نمونه روش بهسازی گفتار تک‌کاناله با این فرض تشریح می‌شود که گفتار از هر دو مدل قطعی و تصادفی پیروی می‌کند [۱۳].

۳-۲-۱- تابع چگالی احتمال ضرایب تبدیل فوریه زمان کوتاه سیگنال نویزی در مدل های تصادفی و قطعی

مدل تصادفی گاوسی: تحت مدل تصادفی و استفاده از این فرض که ضرایب تبدیل فوریه گسسته سیگنال تمیز و نویز هر دو توزیع مختلط گاوسی با میانگین صفر دارند، قسمت حقیقی / موهومی ضرایب تبدیل فوریه گسسته سیگنال نویزی دارای توزیع گاوسی با میانگین صفر رابطهٔ (۱۳) است که برای سادگی، رابطه صرفاً برای قسمت حقیقی نوشته شده است.

$$P_{Y_R|ST}\left(y_R(k,i)|\mathrm{st}\right)=\frac{1}{\sqrt{2\pi\times\sigma^2_{Y_R}(k,i)}}\times\exp\left\{-\frac{|y_R(k,i)|^2}{2\times\sigma^2_{Y_R}(k,i)}\right\}, \quad (۱۳)$$

در این رابطه، ST نشان دهندهٔ سیگنال گفتار مبتنی بر مدل تصادفی است. همچنین، $\sigma^2_{Y_R}(k,i)$ واریانس قسمت حقیقی ضرایب تبدیل فوریه گسسته سیگنال نویزی و برابر با جمع واریانس نویز و واریانس سیگنال تمیز است.

$$\sigma^2_{Y_R}(k,i)=\sigma^2_{S_R}(k,i)+\sigma^2_{W_R}(k,i). \quad (۱٤)$$

مدل قطعی: تحت مدل قطعی گفتار فرض می‌شود $Y(k,i)$ جمع متغیر قطعی $S(k,i)$ و متغیر تصادفی $W(k,i)$ است؛ بنابراین، با فرض توزیع گاوسی مختلط با میانگین صفر برای ضرایب تبدیل فوریه گسسته نویز، ضرایب تبدیل فوریه گسسته سیگنال نویزی، توزیع گاوسی با میانگین غیرصفر خواهد داشت.

$$P_{Y_R|D}\left(y_R(k,i)|\mathrm{d}\right)=\frac{1}{\sqrt{2\pi\times\sigma^2_{W_R}(k,i)}}\times\exp\left\{-\frac{\left|y_R(k,i)-E\left[Y_R(k,i)\right]\right|^2}{2\times\sigma^2_{W_R}(k,i)}\right\}, \quad (۱۵)$$

در این رابطه، D نشان‌دهندهٔ قطعی‌بودن مدل گفتار است. در این صورت $E\left[Y_R(k,i)\right]=s_R(k,i)$ و $E\left[Y_R(k,i)\right]=\mathrm{Re}\left\{E\left[Y(k,i)\right]\right\}$ است.

۳-۲-۲- تخمینگر کمترین میانگین مربعات خطا

مدل تصادفی گاوسی: تخمین سیگنال تمیز مطابق بخش ۳-۱-۱ است.

مدل قطعی: تحت مدل قطعی گفتار، ضرایب تبدیل فوریه گفتار تمیز، معین اما ناشناخته فرض می‌شوند. به این معنی که $P_S\left(s(k,i)\right)=\delta\left(s(k,i)-s'(k,i)\right)$ و $s'(k,i)$ مقدار ضریب تبدیل فوریه گسسته سیگنال تمیز، معین و $\delta(.)$ تابع ضربه است. پس تخمینگر کمترین میانگین مربعات خطا مطابق با رابطهٔ (۱٦) خواهد بود:

$$\hat{s}(k,i)=E\left[S(k,i)\,|\,y(k,i)\right]=s'(k,i) \quad (۱٦)$$

در این رابطه، $s'(k,i)=E\left[Y(k,i)\right]$ است.

۳-۲-۳- تخمین $E\left[Y(k,i)\right]$ با استفاده از مدل قطعی

فرض می‌شود سیگنال تمیز گفتار با جمع P تابع نمایی نزولی با فرکانس ثابت بازنمایی می‌شود:

$$s_t(n)=\sum_{p=1}^{P}a_p e^{j\phi_p}e^{\left(-d_p+jv_p\right)n} \quad (۱۷)$$

$s_t(.)$ نمونه‌ها در حوزهٔ زمان، n اندیس نمونهٔ زمانی، a_p دامنه، ϕ_p فاز، d_p ضریب میرایی و v_p مؤلفهٔ فرکانس pام است؛ درنتیجه، ضرایب تبدیل فوریه گسسته در هر بین فرکانسی k با جمع P تابع نمایی مختلط توصیف می‌شوند [۱۳]. بعد از شیفت و پنجره‌گذاری $s_t(n)$ رابطهٔ (۱۸) به دست می‌آید:

$$s_w(n-mM)=\sum_{p=1}^{P}a_p e^{j\phi_p}e^{\left(-d_p+jv_p\right)(n-mM)}w(n), \quad (۱۸)$$

$w(n)$ پنجرهٔ آنالیز به طول K و M مقدار شیفت هر فریم $(M\le K)$ است. اکنون از سیگنال پنجره‌شده $s_w(n)$ تبدیل فوریه گسسته گرفته می‌شود:

$$s(k,\mathrm{m})=\sum_{n=0}^{K-1}\sum_{p=1}^{P}a_p e^{j\phi_p}e^{\left(-d_p+jv_p\right)(n-mM)}\times w(n)e^{-j\omega_k n}=s(k,0)e^{-\left(-d_p+jv_p\right)(mM)}, \quad (۱۹)$$

$\omega_k=2\pi/L$، L اندازهٔ تبدیل فوریه گسسته $\left(L\ge K\right)$ و m اندیس فریم است. رابطهٔ (۱۹) به فرم $s(k,\mathrm{i})=z^m s(k,0)$ نوشته می شود؛ با این فرض که $z=e^{\left(d_p-jv_p\right)M}$ است. حال اگر نویز در فریم‌های $m=i-n_1,\ldots,i+n_2$ ایستان باشد و M به اندازهٔ کافی بزرگ انتخاب شود، نویز در بازهٔ $m=i-n_1,\ldots,i+n_2$ سفید خواهد بود [۱۳].

برای تخمین d_p و v_p از روش ماتریس پنسِل استفاده می‌شود که در بخش قبل، بهترین روش تخمین

انتخاب شد. تخمین سیگنال گفتار تمیز در بین فرکانسی k و فریم i به‌صورت رابطهٔ (۲۰) است:

$$\hat{s}(k,\mathrm{i}) = E\left[Y(k,i)\right] \approx \frac{1}{(n_2+n_1+1)} \times \sum_{m=i-n_1}^{i+n_2} y(k,\mathrm{m}) e^{\left(-d_p+jv_p\right)(i-m)M}. \quad (۲۰)$$

۳-۳- تخمین کمترین میانگین مربعات خطا تحت مدل قطعی - تصادفی گفتار

برای پیداکردن تخمینگر کمترین میانگین مربعات خطا تحت یک مدل ترکیبی قطعی - تصادفی گفتار، یک مدل کاملاً عمومی استفاده می‌شود که در آن تصمیم‌گیری براساس احتمال بین مدل قطعی و تصادفی صورت می‌گیرد و عدم قطعیت حضور گفتار نیز در نظر گرفته شده است. در این مدل ابتدا مجموعه $\alpha = \{A, D, ST\}$ معرفی می‌شود. $\alpha = A$، $\alpha = D$ و $\alpha = ST$ به‌ترتیب نشان‌دهندهٔ حضورنداشتن گفتار، تولید گفتار با مدل قطعی و تولید گفتار با استفاده از مدل تصادفی است. تمام احتمالات در این بخش در بین فرکانسی k و فریم i صادق است که برای سادگی از نوشتن آن صرف نظر شده است.

برای پیداکردن تخمینگر بهینهٔ کمترین میانگین مربعات خطا، ابتدا عبارت شرطی $E[S \mid y]$ محاسبه می‌شود:

$$\hat{s} = E[S \mid y] = E[S_R \mid y_R] + jE[S_I \mid y_I] = \hat{s}_R + j\hat{s}_I, \quad (۲۱)$$

$$\hat{s}_R = E[S_R \mid y_R, d] P_{D|Y_R}(d \mid y_R) + E[S_R \mid y_R, st] P_{ST|Y_R}(st \mid y_R). \quad (۲۲)$$

زمانی که $\alpha = A$ است، $s_R = 0$ خواهد بود. احتمال‌های شرطی $P_{D|Y_R}(d \mid y_R)$ و $P_{ST|Y_R}(st \mid y_R)$ طبق قضیه بیز به‌صورت زیر محاسبه می‌شوند:

$$P_{D|Y_R}(d \mid y_R) = \frac{\Lambda_{DR}}{\Lambda_{DR}+\Lambda_{STR}+1}, \quad (۲۳)$$

$$P_{ST|Y_R}(st \mid y_R) = \frac{\Lambda_{STR}}{\Lambda_{DR}+\Lambda_{STR}+1}. \quad (۲۴)$$

با فرض اینکه

$$\Lambda_{DR} = \frac{P_{Y_R|D}(y_R \mid d) P_D(d)}{P_{Y_R|A}(y_R \mid a) P_A(a)}, \quad (۲۵)$$

$$\Lambda_{STR} = \frac{P_{Y_R|ST}(y_R \mid st) P_{ST}(st)}{P_{Y_R|A}(y_R \mid a) P_A(a)}, \quad (۲۶)$$

جدول (۱): مقادیر احتمال‌های پیشین طبق مرجع [۱۳]

نوع احتمال پیشین	مقدار احتمال پیشین
$P_D(d)$	۰/۰۲۱
$P_{ST}(st)$	۰/۲۲
$P_A(a)$	۰/۷۵۹

در این معادلات، احتمال‌های $P_{ST}(st), P_D(d)$ و $P_A(a)$ به‌ترتیب نشان دهندهٔ احتمال های پیشین در زمانی است که بین فرکانسی گفتار معین، بین فرکانسی گفتار تصادفی و بین فرکانسی گفتار غایب اند (سکوت). برای محاسبهٔ این احتمال‌ها فرض می‌شود برای یک گفتار انگلیسی متوسط دورهٔ گفتار مصوت ۷۸٪ از زمان است، فرکانس اساسی گفتار f_0 نیز بین ۵۰ و ۵۰۰ هرتز است و برای بیشتر صداهای مصوت گفتار انرژی گفتار عمدتاً تا حدود $f_c = 2000$ Hz خواهد بود. حال احتمال ها مطابق رابطه‌های (۲۷)، (۲۸) و (۲۹) محاسبه می‌شوند:

$$P_D(d) = 0.78 \times \frac{f_c}{f_0} \Big/ \frac{K}{2}, \quad (۲۷)$$

$$P_{ST}(st) = 0.22, \quad (۲۸)$$

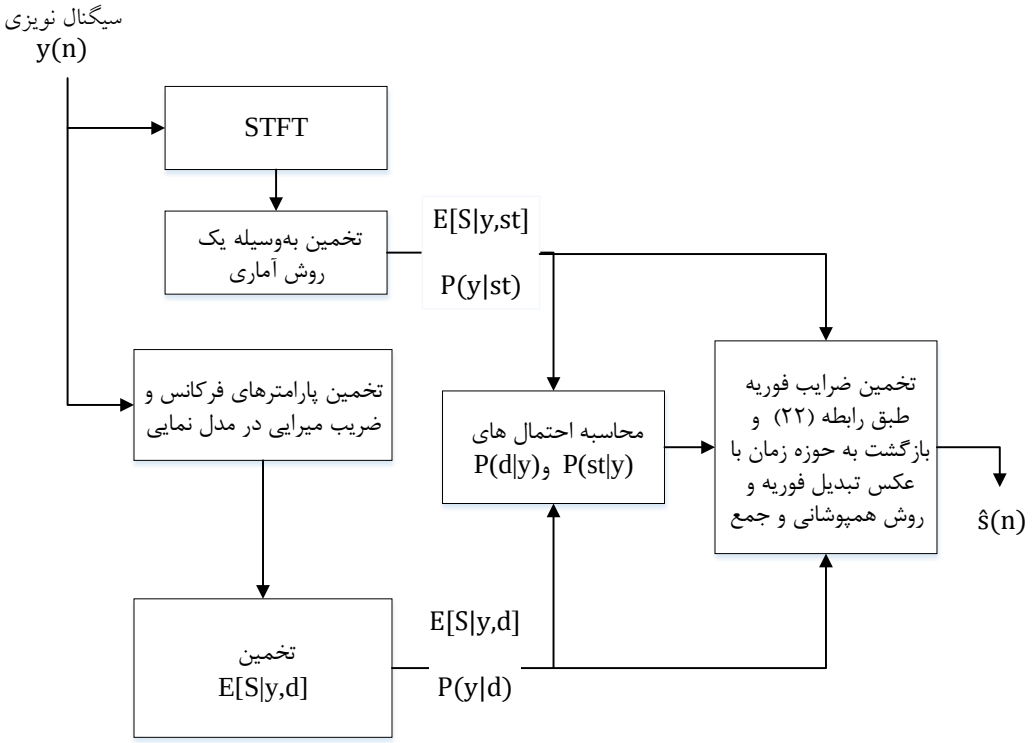
$$P_A(a) = 1 - P_D(d) - P_{ST}(st), \quad (۲۹)$$

K طول فریم است. برای فرکانس نمونه برداری ۱۶ کیلوهرتز، طول فریم ۴۸۰ نمونه و فرکانس اساسی ۳۰۰ هرتز مقادیر احتمال پیشین طبق جدول (۱) به دست می آیند

[۱۳]. در روابط (۲۵) و (۲۶) احتمال $P_{Y_R|A}(y_R\,|\,a)$ مطابق با رابطهٔ (۳۰) است:

$$P_{Y_R|A}(y_R\,|\,a)=\frac{1}{\sqrt{2\pi\sigma_{W_R}^2}}\exp\left\{-\frac{(y_R)^2}{2\sigma_{W_R}^2}\right\}. \qquad (۳۰)$$

بلوک دیاگرام روش بهسازی به روش ترکیب مدل قطعی و تصادفی در شکل (۳) آمده است.



شکل (۳): بلوک دیاگرام بهسازی گفتار با روش قطعی - تصادفی

٤- روش پیشنهادی بهسازی با ترکیب مدل نمایی و مدل تصادفی *t* location-scale

در این مقاله، از توزیع تصادفی *t* location-scale به‌جای توزیع تصادفی گاوسی در ترکیب با مدل قطعی نمایی استفاده شده است. در این حالت، با توجه به گاوسی‌بودن نویز، روابط (۱۵) و (۳۰) به قوت خود باقی خواهند ماند و فقط رابطهٔ $P_{Y|ST}(.)$ با فرض توزیع جدید تغییر می‌کند که مطابق با رابطهٔ (۳۱) استخراج شده است:

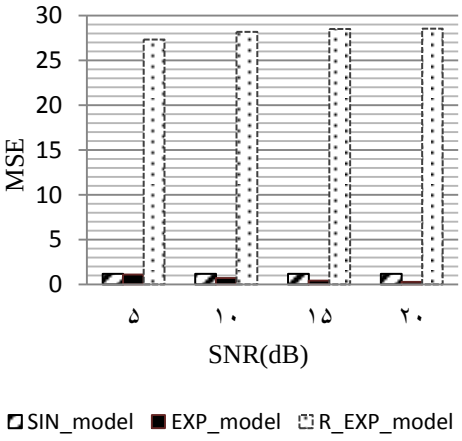
$$P_{Y_R|ST}\left(y_R(k,i)\,|\,st\right)=\frac{\exp\left(-\frac{\varsigma_R}{2}\right)\left(\frac{\xi_R(\nu-2)}{2}\right)^{(\nu+1)/2}}{\sigma_{W_R}\sqrt{\xi_R\pi(\nu-2)}\Gamma\left(\frac{\nu}{2}\right)}\times \qquad (۳۱)$$
$$\sum_{j=0}^{\infty}\frac{1}{j!}\left(\frac{\varsigma_R}{2}\right)^j\Gamma\left(\frac{\nu+1}{2}\right)\psi\left(\frac{\nu+1}{2},\frac{\nu}{2}+1-j;\frac{\xi_R(\nu-2)}{2}\right).$$

همچنین، برای توصیف $E\left[Y_R(k,i)\right]$ در رابطهٔ (۱۵)، مدل نمایی تعمیم داده شده در بخش ۳ به کار رفته است؛ با این تفاوت که از روش ماتریس پِنسِل برای تخمین فرکانس و ضریب میرایی استفاده شده و P نیز برابر ٤٠ در نظر گرفته شده است. با این فرضیات، سیگنال گفتار در مدل قطعی مطابق رابطهٔ (۳۲) تخمین زده می‌شود:

$$\hat{s}(k,\mathrm{i}) = E\left[Y(k,i)\right] \approx \frac{1}{P(n_2+n_1+1)} \times \sum_{m=i-n_1}^{i+n_2} \sum_{i=1}^{P} y(k,\mathrm{m}) e^{(-d_p + jv_p)(i-m)M}. \quad (۳۲)$$

۵- شبیه‌سازی و نتایج

در این بخش، ابتدا بهترین مدل قطعی برای ترکیب با مدل تصادفی در بهسازی گفتار بررسی شده است و سپس بهترین مدل قطعی انتخاب‌شده در ترکیب با تخمینگر کمینهٔ میانگین مربعات خطا مبتنی بر مدل تصادفی t location-scale برای بهسازی گفتار به کار می‌رود.



شکل (٤): خطای تخمین $E\left[Y(k,i)\right]$ در مدل‌های قطعی مختلف.

۵-۱- مقایسهٔ مدل‌های قطعی مختلف در بهسازی گفتار قطعی - تصادفی

هدف در این بخش، پیداکردن بهترین مدل قطعی برای استفاده در روش بهسازی گفتار قطعی - تصادفی است. مدل قطعی $E\left[Y(k,i)\right]$ را تخمین می زند که با فرض سیگنال گفتار قطعی، $E\left[W(k,i)\right]=0$ و طبق رابطهٔ (۱)، $E\left[Y(k,i)\right]=E\left[S(k,i)\right]$ خواهد بود. برای مقایسهٔ بهترین مدل، از رابطهٔ (۳۳) استفاده می شود که میانگین زمانی مربعات خطای تخمین تبدیل فوریه زمان کوتاه سیگنال گفتار تمیز را محاسبه می‌کند:

$$MSE = \frac{\sum_{i=1}^{Nframe}\left(\sum_{k=1}^{K}\left|E\left[Y(k,i)\right]-S(k,i)\right|^2\right)}{Nframe}, \quad (۳۳)$$

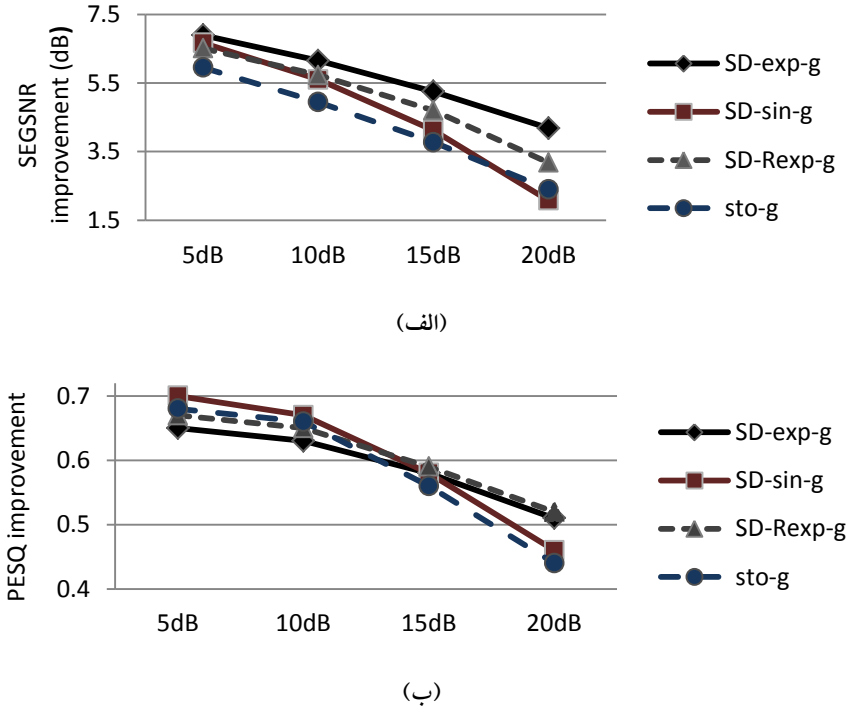
K تعداد بین فرکانسی، $Nframe$ تعداد فریم‌ها و $S(k,i)$ تبدیل فوریه زمان کوتاه سیگنال گفتار تمیز است. سه مدل قطعی بررسی‌شده در این مقاله، شامل مدل سینوسی (SIN_model)، مدل نمایی با پارامتر $P=1$ (R_EXP_model) و مدل نمایی با پارامتر $P=40$ (EXP_model) هستند.

در نخستین شبیه‌سازی، سیگنال گفتار یک دقیقه‌ای از دیتابیس TIMIT با فرکانس نمونه برداری ١٦ کیلوهرتز به کار رفته است [۲۲]. همچنین، طول فریم ٤٨٠ نمونه ای با فریم شیفت ۵۰٪، تبدیل فوریه ۲۰٤۸ نقطه ای و نویز سفید در سیگنال به نویزهای مختلف استفاده شده است. مقدار M در رابطهٔ (۳۲) نصف طول فریم، $n_1=1$، $n_2=0$ قرار داده شده است و v_p و d_p نیز از الگوریتم ماتریس پنسِل تخمین زده می شوند. همان‌طور که در شکل (٤) مشخص است، مدل پیشنهادی نمایی با پارامتر $P=40$، به لحاظ کمترین میانگین مربعات خطای تخمین، بهترین نتیجه را دارد.

۵-۲- مقایسهٔ روش‌های بهسازی گفتار با مدل‌های ترکیبی قطعی - تصادفی و مدل‌های تصادفی

در این بخش، الگوریتم بهسازی گفتار تحت مدل قطعی - تصادفی با سایر روش‌های بهسازی معمول مقایسه می شود که متکی به مدل تصادفی اند. برای ارزیابی نیز از معیارهای اندازه‌گیری ارزیابی ادراکی کیفیت گفتار[۱۳] (PESQ) و نسبت سیگنال به نویز قطعه ای[۱٤] (segSNR) طبق پارامترهای گفته‌شده در [۱۰] استفاده می شود. داده‌های گفتاری بررسی شده در آزمایشات حدود شش دقیقه سیگنال

از مجموعه دادگان TIMIT هستند که به صورت تصادفی از این پایگاه داده انتخاب شده‌اند و مشتمل بر ۵۶ فایل صوتی از گویندگان مرد و ۵۸ فایل صوتی از گویندگان خانم است [۲۲]. فرکانس نمونه‌برداری ۱۶ کیلوهرتز، طول فریم ۴۸۰ نمونه با فریم شیفت ۵۰٪، تبدیل فوریه ۲۰۴۸ نقطه و نویزهای white، pink، volvo، F16، m109 و HF-channel از مجموعه دادگان نویز Noisex-92 برای آزمایش‌ها انتخاب شده‌اند [۲۳]. مقدار احتمال‌های اولیهٔ $P_{ST}(st)$، $P_D(d)$ و $P_A(a)$ مطابق با جدول (۱) مقداردهی می‌شوند. در رابطهٔ (۳۲) مقدار M برابر نصف طول فریم، $n_1=1$، $n_2=0$ قرار داده شده است و ν_p و d_p نیز از الگوریتم ماتریس پنسیل تخمین زده می‌شوند. با توجه به اینکه در نسبت سیگنال به نویزهای پیشین پایین، تخمین فرکانس و ضریب میرایی از دقت کافی برخوردار نیست، در حالت $\xi(k,i) \le -7dB$ صرفاً از مدل تصادفی در بهسازی استفاده می‌شود [۱۳].



شکل (۵): مقایسهٔ عملکرد مدل تصادفی گاوسی و مدل ترکیبی گاوسی / قطعی برای بهسازی در حضور نویز سفید (الف) بهبود معیار segSNR نسبت به حالت نویزی، (ب) بهبود معیار PESQ نسبت به حالت نویزی.

همچنین، در تخمین $\xi(k,i)$، ضریب هموارکنندگی $\alpha=0.98$ و $\tilde{\xi}(k,i)=\min\left(\tilde{\xi}(k,i),-15\ \text{dB}\right)$ در نظر گرفته می شود. گفتنی است پارامترهای طول فریم، فریم شیفت و M، طبق مرجع [۱۳] و پارامترهای P، n_1 و n_2 نیز با سعی و خطا برای دستیابی به عملکرد مناسب در بهسازی انتخاب شده‌اند.

روش های بهسازی گفتار مختلف در این مقاله با علائم اختصاری زیر معرفی می‌شوند. بهسازی گفتار تحت مدل تصادفی گاوسی با sto-g [۲]، بهسازی گفتار تحت مدل تصادفی location-scale t با sto-tls [۱۰]، بهسازی گفتار تحت مدل گاوسی – سینوسی با SD-sin-g، بهسازی گفتار تحت مدل گاوسی – نمایی هندریکس با پارامتر $P=1$ با SD-Rexp-g [۱۳]، بهسازی گفتار تحت مدل گاوسی – نمایی تعمیم‌یافته با $P=40$ با SD-exp-g (روش

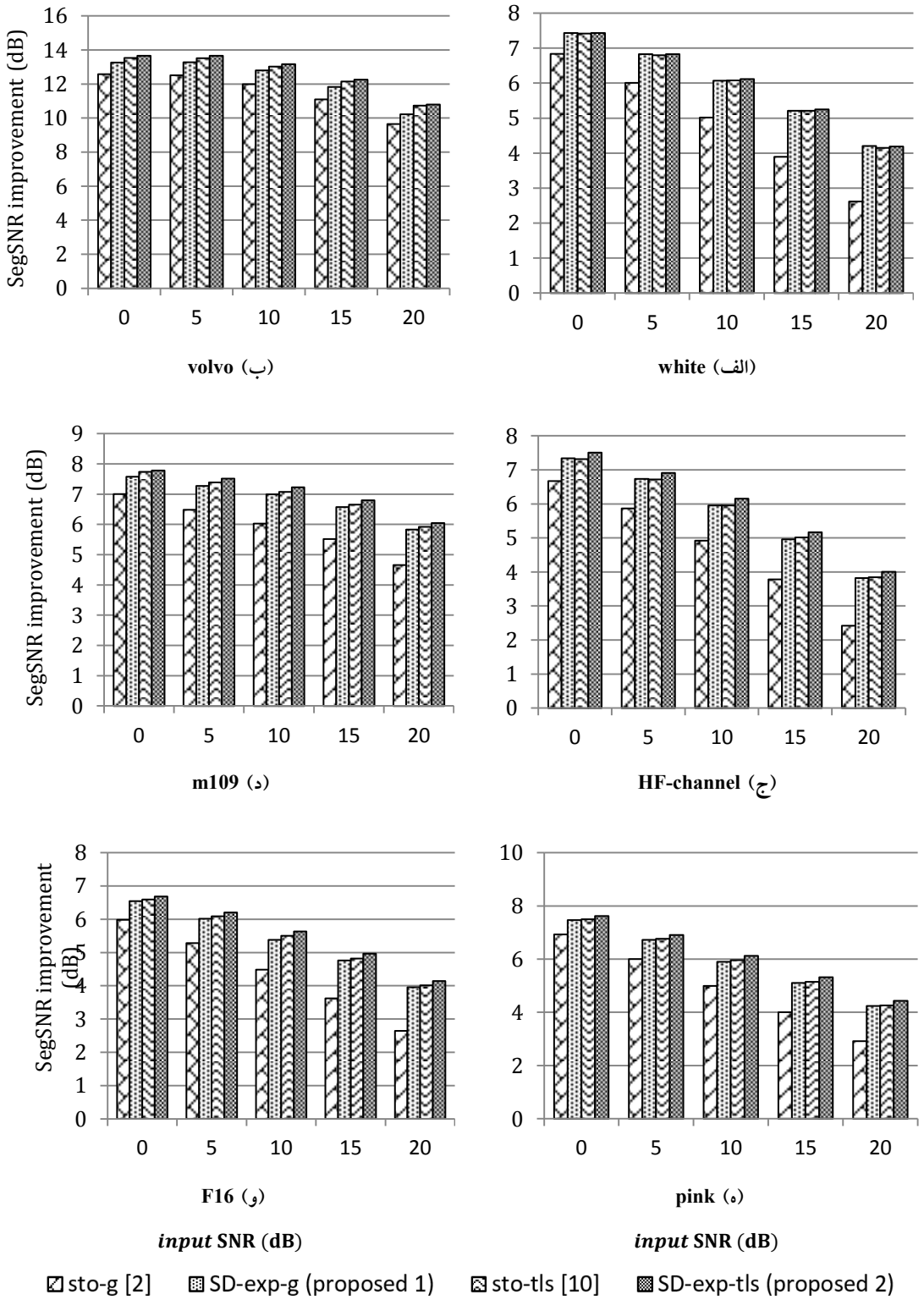
پیشنهادی نخست) و درنهایت، بهسازی گفتار با مدل t location-scale – نمایی با پارامتر $P = 40$ با SD-exp-tls (روش پیشنهادی دوم) نمایش داده می‌شوند.

۵-۲-۱- مقایسهٔ روش‌های بهسازی معرفی‌شده تحت نویز سفید

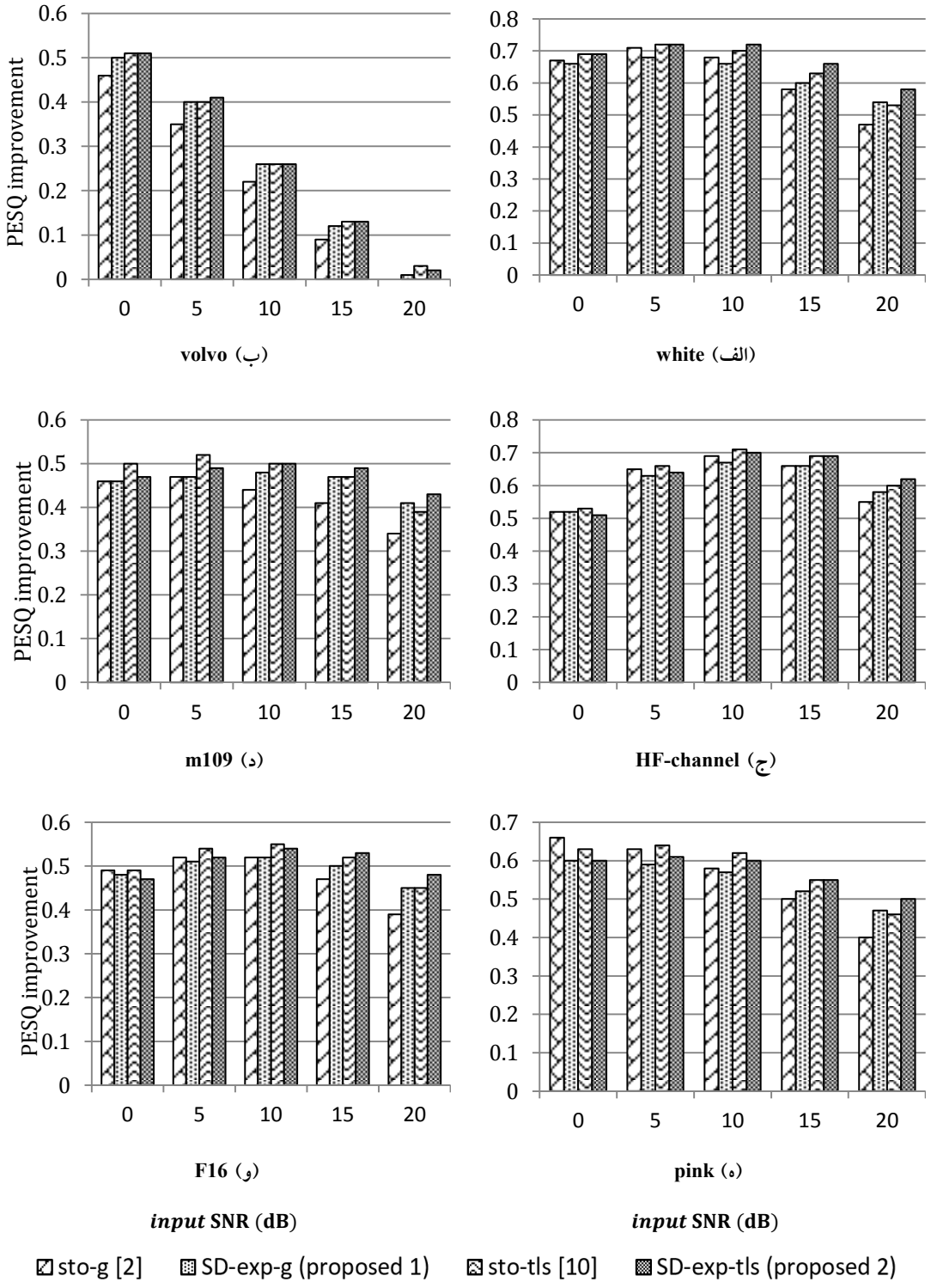
در شکل (۵) عملکرد الگوریتم‌های SD-exp-g، SD-sin-g، SD-Rexp-g و sto-g با هم مقایسه می‌شوند که همگی بر مبنای مدل تصادفی گاوسی‌اند. در این شکل، بهبود معیارهای segSNR و PESQ در حالتی بررسی می‌شود که سیگنال گفتار شش دقیقه‌ای با نویز سفید در سیگنال به نویز ورودی در رنج ۵ تا ۲۰ دسی بل آلوده شده است.

طبق شکل (۵.الف)، ترکیب مدل نمایی تعمیم یافتهٔ پیشنهادی با مدل تصادفی گاوسی، روش SD-exp-g، نسبت به روش بهسازی SD-Rexp-g به بهبود ۰/۳ دسی بلی در معیار segSNR منجر شده است. طبق شکل (۵.ب)، الگوریتم SD-sin-g، بهبود ۰/۰۲ در معیار PESQ نسبت به الگوریتم SD-Rexp-g دارد. با توجه به اینکه معیار segSNR و PESQ را نمی توان هم‌زمان با هم بهتر کرد و همیشه یک چالش در بهبود هم‌زمان این دو معیار وجود دارد و البته بهبود معیار PESQ در الگوریتم SD-sin-g کمتر از ۰/۰۳ است، در مجموع، روش SD-exp-g روش بهتری نسبت به دو روش دیگر ارزیابی می‌شود. در آزمایش بعد، مدل قطعی که به کار می‌رود، مدل نمایی تعمیم یافتهٔ پیشنهادی است که ترکیب آن با دو مدل تصادفی گاوسی و t location-scale بررسی می‌شود.

در شکل (٦) و (۷)، عملکرد الگوریتم SD-exp-tls (روش پیشنهادی دوم) به‌ازای $\upsilon = 3.1$، با SD-exp-g (روش پیشنهادی نخست)، sto-g [۲] و sto-tls [۱۰] مقایسه می شود. معیار مقایسهٔ الگوریتم ها، متوسط بهبود معیار segSNR و PESQ در شش نویز مختلف و در رنج SNRهای صفر تا ۲۰ دسی بِل است. مطابق با شکل (٦)، استفاده از مدل نمایی تعمیم‌یافتهٔ پیشنهادی در ترکیب با مدل تصادفی گاوسی، یعنی روش SD-exp-g، به بهبود چشمگیری در معیار سیگنال به نویز قطعه‌ای نسبت به مدل sto-g منجر شده است که اهمیت ترکیب یک مدل قطعی مناسب با مدل تصادفی را در افزایش بازدهی بهسازی گفتار نشان می دهد. روش sto-tls صرفاً مبتنی بر یک مدل تصادفی است و از توزیع منطبق‌تری با داده های گفتار تمیز بهره می جوید؛ بنابراین، دقت درخور قیاسی با روش SD-exp-g دارد و البته ترکیب مدل قطعی نمایی با آن، طبق روش پیشنهادی SD-exp-tls در برخی نویزها به بهبود در معیار سیگنال به نویز قطعه‌ای منجر شده است. در شکل (۷) بهبودی معیار PESQ ارزیابی شده است که در مجموع، تفاوت چشمگیری بین انواع روش های بهسازی مشهود نیست و از این لحاظ روش ها می توانند با یکدیگر مقایسه شوند.



شکل (٦): بهبود معیار segSNR نسبت به حالت نویزی در روش مدل تصادفی گاوسی (sto-g)، مدل ترکیبی گاوسی - نمایی (SD-exp-g)، مدل تصادفی *t* location-scale (sto-tls) و مدل ترکیبی *t* location-scale – نمایی (SD-exp-tls) برای بهسازی در حضور شش نویز ایستان از مجموعه داده‌های نویز Noisex-92.



شکل (۷): بهبود معیار PESQ نسبت به حالت نویزی در روش مدل تصادفی گاوسی (sto-g)، مدل ترکیبی گاوسی - نمایی (SD-exp-g)، مدل تصادفی *t* location-scale (sto-tls) و مدل ترکیبی *t* location-scale – نمایی (SD-exp-tls) برای بهسازی در حضور شش نویز ایستان از مجموعه داده‌های نویز Noisex-92.

در جدول (۲) میزان کارایی متوسط چهار روش بهسازی نمایش داده شده که در شش نویز مختلف و پنج مقدار سیگنال به نویز به دست آمده است. مطابق با این جدول، روش ترکیبی گاوسی – نمایی (روش پیشنهادی نخست) در مجموع، موجب بهبود چشمگیری در حدود ۰/۹ دسی‌بل در معیار سیگنال به نویز قطعه‌ای شده است؛ البته بهبود حدود ۰/۱ دسی‌بلی روش ترکیب مدل نمایی با مدل تصادفی tls (روش پیشنهادی دوم) نسبت به مدل تصادفی صرف نیز مشهود است.

جدول (۲): مقایسهٔ عملکرد میانگین الگوریتم‌های sto-g، sto-tls، SD-exp-g و SD-exp-tls در حضور شش نویز مختلف از دادگان Noisex-92 و پنج مقدار سیگنال به نویز

الگوریتم	متوسط بهبود معیار segSNR (dB)	متوسط بهبود معیار PESQ
sto-g	٦/٠٨	٠/٤٨٦
SD-exp-g (proposed 1)	٧/٠١	٠/٤٩٩
sto-tls	٧/٠٩	٠/٥٢
SD-exp-tls (proposed 2)	**٧/٢١**	**٠/٥٢١**

جدول (۳): رتبهٔ میانگین الگوریتم‌های sto-g، sto-tls، SD-exp-g و SD-exp-tls در حضور شش نویز مختلف از دادگان Noisex-92 و پنج مقدار سیگنال به نویز

الگوریتم	معیار segSNR	معیار PESQ
sto-g	٣/٨٦	٢/٨٣
SD-exp-g (proposed 1)	٢/٦٦	٢/٦٦
sto-tls	٢/١	**١/٤٦**
SD-exp-tls (proposed 2)	**١/٠٣**	١/٥٦

تفاوت در معیار PESQ به‌طور متوسط حدود ۰/۰۱ تا ۰/۰۳ بوده است که از این لحاظ، روش ها تفاوت معناداری ندارند. برای بررسی علمی‌تر، تست فریدمن[15] روی رتبهٔ چهار روش بهسازی در ۳۰ شرایط مختلف آزمایش، یعنی شش سیگنال نویز مختلف و پنج مقدار سیگنال به نویز اجرا شده است. رتبهٔ متوسط روش‌ها در هر دو معیار segSNR و PESQ در جدول (۳) نشان داده شده است. در تست فریدمن که با استفاده از تابع $p = \text{friedman}(.)$ در نرم‌افزار MATLAB روی رتبهٔ روش‌ها در نتایج سیگنال به نویز قطعه‌ای اجرا شده، مقدار p برابر با $1/7\times10^{-17}$ نشان‌دهندهٔ تفاوت معنادار چهار روش به لحاظ آماری است. همچنین، برای تأیید مؤثربودن ترکیب مدل نمایی با مدل تصادفی تست فریدمن روی دوبه‌دوی روش ها انجام شده و مقدار p برابر با $4/3\times10^{-8}$ در مقایسهٔ دو روش SD-exp-g و sto-g و نیز $4/3\times10^{-8}$ در مقایسهٔ دو روش SD-exp-tls و sto-tls حاصل شده است. مقدار p کمتر از ۰/۰۱ نشان دهندهٔ تفاوت معنادار روش های بهسازی ارزیابی‌شده و مؤثربودن ترکیب مدل نمایی تعمیم‌یافتهٔ پیشنهادی در هر دو مدل تصادفی است. همچنین، انجام تست فریدمن روی دو مدل SD-exp-tls و SD-exp-g و حصول مقدار p برابر با $8/9\times10^{-7}$ کارایی بالاتر مدل تصادفی t location-scale را در مقابل مدل گاوسی در ترکیب با مدل نمایی پیشنهادی نشان می دهد. انجام تست فریدمن در معیار PESQ و مقدار p به دست‌آمده، تفاوت آماری معناداری را بین دوبه‌دوی روش ها به روال بالا نشان نداد؛ به این ترتیب، برابری نسبی میانگین رتبه های عملکرد دو روش sto-tls و SD-exp-tls (روش پیشنهادی دوم) و نیز sto-g و SD-exp-g (روش پیشنهادی نخست) طبق جدول (۳)، کارایی نداشتن مدل ترکیبی قطعی – نمایی را در قیاس با مدل تصادفی صرف در بهبود معیار PESQ نشان می‌دهد.

٦- نتیجه‌گیری

در این مقاله، یک روش نوین بهسازی گفتار در حالت تک کاناله با استفاده از ترکیب مدل قطعی نمایی و مدل تصادفی ارائه شد. روش پیشنهادشده، تعمیمی بر روش معرفی‌شدهٔ هندریکس و همکارانش در سال ۲۰۰۷ است که از توزیع تصادفی گاوسی و لاپلاس به‌عنوان مدل تصادفی و از مدل نمایی به‌عنوان مدل قطعی استفاده کردند. در این مقاله از توزیع جدید t location-scale به‌عنوان مدل تصادفی استفاده شد و همچنین، مدل نمایی استفاده‌شده در مقالهٔ هندریکس با افزایش پارامتر P و تغییر روش تخمین فرکانس از اسپریت[16] به ماتریس پنسیل بهبود داده شد. نتایج پیاده سازی در شش نویز مختلف نشان داد روش جدید ارائه‌شده، یعنی ترکیب مدل نمایی

تعمیم‌یافته و مدل تصـادفی t location-scale، بـه بهبـود معیار segSNR می توانـد منجـر شـود و کـارایی درخـور قیاسی را در معیار PESQ در مقایسه با روش‌های بهسازی دیگر نتیجه دهد.

مراجع

[1] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", IEEE Trans. on Acoust., Speech, Signal Process, Vol. 32, No. 6, pp. 1109–1121, Dec 1984.

[2] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction", IEEE Trans. Acoust., Speech, Signal Process, Vol. 27, No. 2, pp. 113–120, Apr 1979.

[3] K. Funaki, "Speech enhancement based on iterative Wiener filter using complex speech analysis", 16th European Signal Processing Conference, pp. 1–5, 25-29 Aug 2008.

[4] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", IEEE Trans. Acoustic., Speech, Signal Process., Vol. 33, No. 2, pp. 443–445, May 1985.

[5] R.J. Macaulay, M.L. Malpass, "Speech enhancement using a soft decision noise suppression filter", IEEE Trans. Acoustic., Speech, Signal Process, Vol. 28, No. 2, pp. 137–145, Apr 1980.

[6] T. Lotter, P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model", EURASIP Journal on Advances in Signal Processing, pp. 1110–1126, Dec 2005.

[7] B. Chen, P.C. Loizou, "A Laplacian-based MMSE estimator for speech enhancement", Speech Communication., Vol. 49, No. 2, pp. 134–143, Feb 2007.

[8] R. Martin, "Speech enhancement based on minimum mean-square error estimation and super Gaussian priors", IEEE Trans. Speech, Audio Process., Vol. 13, No. 5, pp. 845–856, Aug 2005.

[9] J.S. Erkelens, R.C. Hendriks, R. Heusdens, et al. , "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, July 2007.

[10] N. Faraji, A. Kohansal, "MMSE and maximum a posteriori estimators for speech enhancement in additive noise assuming a t-location-scale clean speech prior", IET Signal Processing, Vol. 12, No. 4, pp. 532-543, June 2018.

[11] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter", IEEE Trans. Acoust., Speech, Signal Process., Vol. 28, No. 2, pp. 137–145, Apr. 1980.

[12] J. Hardwick, C. Yoo, and J. Lim, "Speech enhancement using the dual excitation speech model", in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Vol. 2, pp. 367–370, 27-30 Apr 1993.

[13] R. Hendriks, R. Heusdens, J. Jensen, "An MMSE estimator for speech enhancement under a combined stochastic-deterministic speech model", IEEE Trans. Audio Speech Lang. Process., Vol. 15, No. 2, pp. 406–415, Jan 2007.

[14] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic+noise model," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Vol. 2, pp. 550–553, 27-30 Apr 1993.

[15] M. C. McCallum and B. J. Guillemin, "Accounting for deterministic noise components in a MMSE STSA speech enhancement framework," in Proc. 12th Int. Symp. Commun. Inf. Technol., pp. 174–179, 2-5 Oct 2012.

[16] M. McCallum, B. Guillemin, "Stochastic-deterministic MMSE STFT speech enhancement with general a priori information", IEEE Trans. Audio, Speech, Lang. Process., Vol. 21, No. 7, pp. 1445–1457, July 2013.

[17] Y. Du, J. Du, L.R. Dai, et al., "'A regression approach to speech enhancement based on deep neural networks", IEEE/ACM Trans. Audio Speech, Lang. Process., Vol. 23, No. 1, pp. 7–19, Jan 2015.

[18] J.G. Proakis, D.G. Manolakis, Digital Signal Processing: Principles, Algorithms and Applications, Prentice Hall, 3rd edition, 1995.

[19] K. Duda, T. P. Zielinski, "Efficiency of the frequency and damping estimation of a real value sinusoid,", IEEE Instrumentation & Measurement Magazine, Vol. 16, No. 2, pp. 48–58, Apr 2013.

[20] T.K. Sarkar, O. Pereira, "Using the Matrix Pencil Method to Estimate the Parameters of a Sum of Complex Exponentials", IEEE Antennas and Propagation Magazine, Vol. 37, No. 1, pp. 48-55, Feb 1995.

[21] T.K. Moon, W.C. Stirling, Mathematical Methods and Algorithms for Signal Processing, Pearson, PAP/CDR edition, 1999.

[22] W.M. Fisher, G.R. Doddington, K.M. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status", in Proceedings of DARPA workshop on speech recognition, pp. 93–99, 1986.

[23] A. Varga, and H.J.M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems", Speech Communication, Vol. 12, No. 3, pp. 247-251, 1993.

[1] Minimum mean squared error

[2] Maximum likelihood
[3] Maximum a posteriori
[4] Super Gaussian
[5] Generalized gamma
[6] Presence probability
[7] Deep Neural Network
[8] Prony method
[9] Steiglitz-McBride method
[10] Matrix pencil
[11] Total least squares
[12] Frequency bin
[13] Perceptual Evaluation of Speech Quality
[14] Segmental Signal to Noise Ratio
[15] Friedman test
[16] ESPRIT