Semnan University

Journal of Modeling in Engineering

Journal homepage: https://modelling.semnan.ac.ir/

ISSN: 2783-2538

Research Article

Speaker Recognition Using Convolutional Neural Network and Neutrosophic

Sadegh Fadaei [a,*], Abdolreza Rashno [b], Abdolsamad Hamidi [b]

[a] *Department of Electrical Engineering, Faculty of Engineering, Yasouj University, Yasouj, Iran*
[b] *Department of Computer Engineering, Engineering Faculty, Lorestan University, Khorramabad, Iran*

PAPER INFO

Paper history:

Received: 16 February 2023
Revised: 19 April 2023
Accepted: 19 July 2023

Keywords:

Spectrogram,
Speaker recognition,
Neutrosophic,
Convolutional neural networks.

ABSTRACT

Speaker recognition is a process of recognizing persons based on their voice which is widely used in many applications. Although many researches have been performed in this domain, there are some challenges that have not been addressed yet. In this research, Neutrosophic (NS) theory and convolutional neural networks (CNN) are used to improve the accuracy of speaker recognition systems. To do this, at first, the spectrogram of the signal is created from the speech signal and then transferred to the NS domain. In the next step, the alpha correction operator is applied repeatedly until reaching constant entropy in subsequent iterations. Finally, a convolutional neural networks architecture is proposed to classify spectrograms in the NS domain. Two datasets TIMIT and Aurora2 are used to evaluate the effectiveness of the proposed method. The precision of the proposed method on two datasets TIMIT and Aurora2 are 93.79% and 95.24%, respectively, demonstrating that the proposed model outperforms competitive models.

DOI: https://doi.org/10.22075/jme.2023.29933.2409

*** Corresponding author.**
E-mail address: s.fadaei@yu.ac.ir

How to cite this article:
Fadaei, S., Rashno, A., & Hamidi, A. (2023). Speaker Recognition Using Convolutional Neural Network and Neutrosophic. Journal of Modeling in Engineering, 21(75), 1-18. doi: 10.22075/jme.2023.29933.2409

مقاله پژوهشی

تشخیص گوینده با شبکه‌های عصبی کانولوشنال و تئوری نتروسافیک

صادق فدایی[۱،*]، عبدالرضا رشنو[۲] ، عبدالصمد حمیدی[۲]

اطلاعات مقاله

دریافت مقاله: ۱۴۰۱/۱۱/۲۷
بازنگری مقاله: ۱۴۰۲/۰۱/۳۰
پذیرش مقاله: ۱۴۰۲/۰۴/۲۸

واژگان کلیدی:
اسپکتروگرام،
تشخیص گوینده،
نتروسافیک،
شبکه‌ی عصبی کانولوشنال.

چکیده

تشخیص گوینده، فرآیند تشخیص افراد بر اساس صوت آنها است که در کاربردهای زیادی مورد استفاده قرار می‌گیرد. اگرچه تاکنون تحقیقات زیادی در زمینه‌ی تشخیص گوینده صورت گرفته است، اما چالش‌هایی وجود دارد که هنوز حل نشده‌اند. در این مقاله به منظور بهبود دقت سیستم های تشخیص گوینده از نتروسافیک و شبکه‌های عصبی کانولوشنال بهره گرفته شده است. در روش پیشنهادی، ابتدا اسپکتروگرام سیگنال صوتی تشکیل می‌گردد سپس اسپکتروگرام به فضای نتروسافیک منتقل می‌شود. در مرحله‌ی بعد عملگرهای بهبود بتا به مجموعه‌های نتروسافیک اعمال می‌شود و این عملیات تا ثابت شدن آنتروپی مجموعه‌های نتروسافیک تکرار می‌گردد. در نهایت یک مدل شبکه‌ی عصبی کانولوشنال برای طبقه‌بندی هیستوگرام پیشنهاد می‌شود. برای ارزیابی و تحلیل روش پیشنهادی از دو پایگاه داده‌ی Aurora2 و TIMIT استفاده شده است. روش پیشنهادی روی پایگاه داده‌ی Aurora2 به دقت ۹۳/۷۹ درصد و روی پایگاه داده‌ی TIMIT به دقت ۹۵/۲۴ درصد دست یافته است که در مقایسه با روش‌های رقیب عملکرد بهتری داشته است.

DOI: https://doi.org/10.22075/jme.2023.29933.2409

۱- مقدمه

صدای انسان دارای ویژگی‌های منحصر به فردی است که برای هر فرد متفاوت بوده و می‌تواند مانند اثر انگشت برای شناسایی افراد به کار گرفته شود. موضوع تشخیص و شناسایی افراد بر اساس سیگنال صوتی آنها از اهمیت ویژه‌ای برخوردار است و در حوزه‌ی پردازش سیگنال‌های صوتی، به تشخیص گوینده معروف است. تاریخچه‌ی تشخیص گوینده به سال ۱۶۶۰ برمی‌گردد اما به صورت رسمی برای اولین بار در سال ۱۹۶۶، یک دادگاه به منظور تشخیص گوینده از طیف سیگنال صوتی استفاده نمود [۱]. چگونگی استخراج ویژگی از سیگنال صوتی تاثیر بسزایی در دقت تشخیص گوینده دارد و تاکنون[۲] روش‌ها و الگوریتم‌های متعددی به منظور استخراج ویژگی‌های سیگنال صوتی ارایه شده است. از جمله‌ی این روش‌ها می‌توان به تبدیل فوریه [۲]، تبدیل ویولت [۳]، فیلترهای گابور [۴] و شبکه‌های عصبی کانولوشنال [۵] اشاره کرد.

یکی از موضوعاتی که اخیراً در حوزه‌ی پردازش سیگنال مطرح شده و به بهبود الگوریتم‌های پردازش سیگنال کمک

* پست الکترونیک نویسنده مسئول: s.fadaei@yu.ac.ir
۱. استادیار، دانشکده مهندسی، دانشگاه یاسوج، یاسوج، ایران
۲. استادیار، دانشکده مهندسی، دانشگاه لرستان، خرم آباد، ایران

استناد به این مقاله:
فدایی, صادق, رشنو, عبدالرضا, و حمیدی, عبدالصمد. (۱۴۰۲). تشخیص گوینده با شبکه های عصبی کانولوشنال و تئوری نتروسافیک. مدل سازی در مهندسی, ۲۱(۷۵), ۱-۱۸. doi: 10.22075/jme.2023.29933.2409

نموده نتروسافیک[۳] است. نتروسافیک شاخه‌ای از علم فلسفه است که به مطالعه‌ی اصل و سرشت پدیده‌های طبیعی و ارتباط آنها با پدیده‌های خیالی می‌پردازد. نتروسافیک، هر موجودیت A و متضاد آن $\mathrm{Anti_A}$ و طبیعت آن $\mathrm{Neut_A}$ را به عنوان یک رویداد در نظر می‌گیرد [۶]. بر اساس این تئوری، هر موجودیت می‌تواند با نمادهای بالا توصیف شود تا بتواند نمایش بهتری از خصوصیات آن ارایه دهد. در حالت پیش فرض، $\mathrm{Neut_A}$ ،A و $\mathrm{Anti_A}$ را می‌توان به صورت مجموعه‌های جدا از هم و بدون اشتراک در نظر گرفت. در بسیاری از موارد، به علت گنگ و نادقیق بودن مرز مجموعه ها در پدیده‌ای که A بر روی آن تعریف می‌شود، این مجموعه‌ها دارای قسمت‌های مشترکی هستند. در سال‌های اخیر تئوری نتروسافیک در کاربرهای متنوعی از پردازش، شامل بخش‌بندی [۶،۷]، آستانه گذاری [۸]، تشخیص لبه در تصویر [۹] و خوشه‌بندی [۱۰] استفاده شده است.

در حالت کلی رویکرد نتروسافیک برای کاربردهای پردازش سیگنال این است که استخراج هرگونه اطلاعات از سیگنال و نیز اعمال هر پردازشی روی آن، توسط سه مجموعه‌ی درستی، نادرستی و عدم قطعیت مدل می‌شود. سپس عملگرهای لازم روی این مجموعه‌ها در جهت برآورده کردن معیارهای آن کاربرد خاص تعریف می‌شوند. در این نگرش، برای هر نوع داده‌ی صوتی یک مجموعه‌ی خاص برای عدم قطعیت داده‌ها تعریف می‌شود. بنابراین، می‌توان هر نوع نویزی در سیگنال را با تعریف مجموعه‌ی عدم قطعیت متناسب با آن تعریف کرد. نوآوری های اصلی این مقاله در موارد زیر خلاصه می شود:

۱- ابتدا اسپکتروگرام سیگنال صوتی به دست می‌آید و سپس حاشیه‌های سفید آن با استفاده از آستانه‌گذاری حذف می‌شوند. علت حذف حاشیه‌ها این است که دارای اطلاعات مفیدی از سیگنال صوتی نیستند.

۲- با استفاده از یک روش جدید، اسپکتروگرام سیگنال صوتی به فضای نتروسافیک انتقال داده می‌شود. انگیزه‌ی اصلی برای این انتقال این است که نویزهای موجود در سیگنال‌های صوتی در فضای نتروسافیک به خوبی مدل می‌شوند و امکان حذف آنها راحت‌تر است.

۳- آنتروپی اسپکتروگرام در فضای نتروسافیک محاسبه می‌شود. این آنتروپی از جمع آنتروپی زیرمجموعه‌های نتروسافیک بدست می‌آید که در آن، توزیع داده¬ها مورد ارزیابی قرار می‌گیرد.

۴- عملگرهای میانگین آلفا و بهبود بتا به اسپکتروگرام در فضای نتروسافیک اعمال می‌شود. این عملگر ها برای مرتبط ساختن زیرمجموعه‌های نتروسافیک به کار می‌روند که در آنها توزیع داده‌ها و آنتروپی تحت تاثیر قرار می‌گیرند.

۵- در نهایت، یک مدل شبکه‌ی عصبی کانولوشن برای دسته‌بندی اسپکتروگرام سیگنال‌های صوتی در فضای نتروسافیک ارائه می‌گردد.

ادامه‌ی مقاله به این صورت سازمان‌دهی شده است: کارهای مرتبط با موضوع تشخیص گوینده در بخش ۲ آمده است. روش پیشنهادی در بخش ۳ معرفی شده است. پایگاه‌های داده‌ی استفاده شده، تنظیم پارامترها در بخش ۴ ارایه شده است. در بخش ۵ به نتایج پیاده‌سازی و بررسی آنها پرداخته شده است. در بخش ۶ محدودیت-های مدل پیشنهادی و مسیر آینده‌ی تحقیق بحث شده و در نهایت در بخش ۷ مقاله جمع‌بندی شده است.

۲-کارهای مرتبط

تاکنون تحقیقات زیادی در زمینه‌ی تشخیص گوینده صورت گرفته است اما هنوز چالش‌هایی در این حوزه وجود دارد که در ادامه به بررسی بخشی از پژوهش‌هایی که در این زمینه صورت گرفته است پرداخته می‌شود. موضوع تشخیص گوینده و روش‌ها و چالش‌های آن در[۱۵-۱۱] به صورت جامع بررسی شده است. در [۱۶]، از تخمین طیف فرکانسی بر اساس تکنیک پنجره‌گذاری چندمخروطی، برای کاهش تاثیر نویز در دقت تشخیص گوینده استفاده شده است. یکی از مشکلات موجود در مساله‌ی تشخیص گوینده، مشکل عدم تطابق است که در [۱۷] با استفاده از الگوریتم های جداسازی سیگنال صوتی دوگوشی ، برای این مشکل چاره‌اندیشی شده است. اگرچه مدل مخلوط گاوسی می تواند برای سیگنال‌های صوتی طولانی‌مدت عملکردی مشابه گوش انسان داشته باشد لکن برای سیگنال‌های صوتی کوتاه‌مدت عملکرد مناسبی ندارد [۱۸]. برای حل این مشکل و بهبود عملکرد الگوریتم‌های مبتنی بر مدل مخلوط گاوسی در مواجهه با سیگنال‌های صوتی کوتاه‌مدت، یک مدل جدید مبتنی بر شبکه‌های عصبی کانولوشنال در [۱۸]

ارایه شده است.

در [۱۹]، به منظور بهبود دقت تشخیص گوینده، یک رهیافت جدید بر اساس یادگیری ماشین معرفی شده است که از یک سنسور صوتی پیزوالکتریک منعطف استفاده می کند. در [۲۰]، به منظور تعیین MFCC[۴]، مجموعه‌ای از الگوریتم‌های پنجره‌گذاری ارایه شده است. با ترکیب مناسب ویژگی‌های LPC[۵] و MFCC، دقت تشخیص گوینده در [۲۱] بهبود داده شده است. در [۲۲]، ابتدا برای تشکیل بردار ویژگی سیگنال صوتی، MFCC استفاده می شود، سپس به کمک SOFM[۶]، طول بردار ویژگی کاهش داده شده و در نهایت با به‌کارگیری یک شبکه‌ی عصبی پرسپترون چندلایه، عملیات تشخیص گوینده پایان می‌پذیرد. در [۲۳]، یک سیستم تشخیص گوینده مبتنی بر ترکیب ویژگی‌های PLP و MFCC و همینطور شبکه‌های عصبی عمیق پیشنهاد شده است.

در [۲۴]، برای بهبود دقت سیستم‌های تشخیص گوینده، شبکه‌ی AM-SincNet[۷] پیشنهاد شده است. این شبکه بر اساس مدل SincNet پیاده‌سازی شده با این تفاوت که از لایه‌ی پیشنهادی AM-Softmax بهره می‌گیرد. در [۲۵]، به منظور تشخیص گوینده، از سیگنال صوتی ویژگی‌های مرتبط استخراج شده و یک مدل بر اساس این ویژگی‌ها ارایه شده است. در [۲۶]، تکنیک آموزش مخالف مبتنی بر PLDA[۸] به کار گرفته شده است که می‌تواند به عنوان یک مدل پنهان برای اصلاح i-vector باشد. در [۵]، تشخیص گوینده با استفاده از یک شبکه‌ی عصبی کانولوشنال انجام می‌گیرد که در آن مولفه‌های SE[۹] و شبکه‌ی عصبی کانولوشنال با هم ترکیب شده و مدل SECNN[۱۰] را تشکیل داده است. در [۲۷]، دقت سیستم تشخیص گوینده توسط یک الگوریتم وزن‌دهی مناسب برای محاسبه‌ی مشخصات آماری Baum-Welch ارایه شده است که این مشخصات آماری به مدل مخلوط گاوسی در i-vector مربوط می‌شود. در [۲۸]، نسبت واریانس برون-کلاسی و درون-کلاسی ویژگی‌ها به کار گرفته شده است که این ویژگی‌ها دربرگیرنده‌ی پیش‌بینی خطی ادراکی[۱۱] و ضرایب کپسترال پیش‌بینی خطی[۱۲] است. یک الگوریتم جدید برای استخراج ویژگی‌های صوتی مبتنی بر فرکانس‌های مدولاسیون زمانی-طیفی در [۲۹] ارایه شده است. در [۳۰]، با به کارگیری عدم قطعیت با استفاده از نتروسافیک، یک روش جدید مبتنی بر شبکه‌های عصبی کانولوشنال پیشنهاد شده است. در [۳۱]، سیستم‌های نگاشت گوینده‌ی i-vector، t-vector و x-vector به منظور بهبود دقت تشخیص گوینده ارتقا داده شده‌اند که در آنها شرایط عدم تطابق ارایه شده توسط NIST[۱۳] و SRE[۱۴] در نظر گرفته شده است. در [۳۲]، با استفاده از مدولاسیون چند-مقیاسی زمانی-طیفی سیگنال صوتی، یک روش جدید برای بازیابی صحبت مبتنی بر محتوا ارایه شده است. در [۳۳]، با نگاشت مناسب بردار ویژگی توسط یک الگوریتم جدید مبتنی بر PCA[۱۵] به نام WCR-PCA[۱۶]، سیستم تشخیص گوینده بهبود داده شده است.

در [۳۴]، به منظور کاهش تعداد پارامترها و در نتیجه کاهش قابل توجه محاسبات، یک شبکه‌ی عصبی کانولوشنال یک بُعدی بر اساس حالت دو بُعدی آن پیشنهاد شده است. لازم به ذکر است که علی‌رغم کاهش قابل توجه حجم محاسبات، دقت سیستم فقط یک درصد افت داشته است. با توجه به اینکه در کاربردهای واقعی، تشخیص گوینده برای سیگنال‌های صوتی کوتاه‌مدت اهمیت زیادی دارد یک مجموعه بردار ویژگی جدید مبتنی بر مدل مخلوط گاوسی برای سیگنال‌های کوتاه‌مدت، در [۳۵] معرفی شده است. در [۳۶]، مکانیزم خود-توجه[۱۷] و شبکه‌ی ResNet با هم ترکیب شده‌اند و تلاش شده است که با تعداد پارامترهای کمتر و در نتیجه محاسبات کمتری، دقت تشخیص گوینده بهبود یابد. در [۳۷]، به منظور تعیین توزیع احتمال صحبت خنثی[۱۸]، مدل مخلوط گاوسی سیگنال صحبت به کار گرفته شده است و بر همین اساس، یک مدل جدید برای کمینه کردن مشکل عدم تطابق بین سیگنال صوتی احساسی و خنثی معرفی شده است. در [۳۸]، با معرفی ویژگی‌های صوتی سیگنال طولانی‌مدت (LTA)[۱۹]، سعی شده نمایش تُنُکی از سیگنال صوتی ارایه

[4] Mel Frequency Cepstral Coefficients (MFCC)
[5] Linear Productive Coding (LPC)
[6] Self-Organizing Feature Map
[7] Additive Margin-SincNet
[8] Probabilistic Linear Discriminant Analysis
[9] Squeeze-and-Excitation (SE)
[10] Squeeze-and-Excitation Convolutional Neural Network
[11] Perceptual Linear Prediction (PLP)
[12] Linear Prediction Cepstral Coefficient (LPCC)
[13] National Institute of Standards and Technology
[14] Speaker Recognition Evaluation
[15] Principal Component Analysis
[16] Weighted-Correlation PCA
[17] Self-attention
[18] Neutral Speech
[19] Long-Term Acoustic

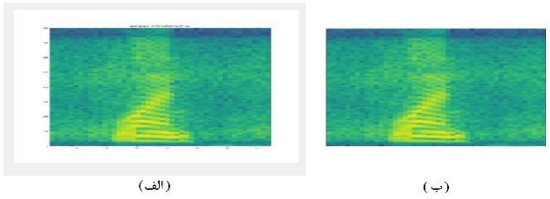
گردد. در [۳۹]، اشاره شده است که شبکه‌های عصبی کانولوشنال برای سیستم‌های متحرک مناسب نیستند و برای شبکه‌ی موبایل[۲۰]، یک روش بر اساس حاشیه‌ی افزودنی پیشنهاد شده است. در [۴۰]، یک روش مقاوم در مقابل نویز مبتنی بر ویولت ارایه شده است که در آن از یک الگوریتم جدید برای محاسبه‌ی آستانه‌های مورد نیاز مولفه-های مختلف ویولت استفاده شده است.

۳- روش پیشنهادی

در این مقاله، به منظور تشخیص گوینده از نتروسافیک و شبکه عصبی کمک گرفته شده است. روش کار به اینصورت است که ابتدا اسپکتروگرام سیگنال صوتی را به دست آورده سپس عملیات پیش‌پردازش روی آن انجام می‌شود. در مرحله‌ی بعد اسپکتروگرام به فضای نتروسافیک منتقل شده و عملگرهای بهبود بتا به مجموعه‌های نتروسافیک اعمال می‌گردد و این عملیات تا ثابت شدن آنتروپی مجموعه‌های نتروسافیک تکرار می‌شود. سپس مجموعه‌های I و T نتروسافیک در یکدیگر ضرب شده و اسپکتروگرام نهایی را تشکیل می‌دهند و در نهایت با اعمال شبکه‌ی عصبی کانولوشنال، گوینده تشخیص داده می‌شود.

۳-۱- تعیین اسپکتروگرام و پیش‌پردازش

ابتدا اسپکتروگرام سیگنال صوتی به دست می‌آید. این ماتریس دارای حاشیه‌های سفید است که این حاشیه‌ها حاوی اطلاعات مفیدی نیستند. از آنجاییکه حاشیه‌ی سفید در اسپکتروگرام دقت شبکه عصبی را پایین می‌آورد بنابراین برای بالا بردن دقت شبکه، مطابق شکل (۱) باید حاشیه‌ی سفید ماتریس حذف شود.



شکل ۱- (الف) اسپکتروگرام سیگنال صوتی و (ب) اسپکتروگرام سیگنال صوتی بعد از پیش‌پردازش.

۳-۲- انتقال اسپکتروگرام به فضای نتروسافیک

در این بخش یک روش جدید در فضای نتروسافیک برای اسپکتروگرام سیگنال‌های صوتی پیشنهاد شده است که نویزهای Street، Airport، Restaurant، TrainStation، Car، Babble و Subway موجود در صوت را کنترل نموده و باعث افزایش دقت سیستم‌های تشخیص گوینده می‌شود. نتروسافیک که با P_{NS} نشان داده می‌شود از سه مجموعه عضویت I، T و F تشکیل شده است. هر داده P با $P(t,i,f)$ توصیف می‌شود. t درصد درست بودن، i درصد عدم قطعیت و f درصد نادرست بودن را نشان می‌دهد. $g(i,j)$ در حوزه‌ی داده‌های اولیه به حوزه‌ی نتروسافیک نگاشت شده و با نماد $P_{NS}(i,j)=\{T(i,j),I(i,j),F(i,j)\}$ نشان داده می شود که $F(i,j)$، $I(i,j)$ و $T(i,j)$ در نتروسافیک کلاسیک به صورت زیر تعریف می‌شوند [۱۰]:

$$T(i,j)=\frac{\bar{g}(i,j)-\bar{g}_{\min}}{\bar{g}_{\max}-\bar{g}_{\min}} \qquad (۱)$$

که در رابطه‌ی بالا $\bar{g}_{\min}$ و $\bar{g}_{\max}$ به ترتیب کمترین و بیشترین مقدار روی ماتریس $\bar{g}$ بوده و $\bar{g}(i,j)$ به صورت زیر تعریف می‌شود [۱۰]:

$$\bar{g}(i,j)=\sum_{m=-\frac{w}{2}}^{\frac{w}{2}}\sum_{n=-\frac{w}{2}}^{\frac{w}{2}} g(i+m,j+n) \qquad (۲)$$

در رابطه‌ی بالا $g(i,j)$ شدت داده در محل (i,j) و $\bar{g}(i,j)$ میانگین محلی $g(i,j)$ در یک پنجره با ابعاد $w\times w$ است. در حقیقت مطابق رابطه‌ی (۱)، ابتدا یک فیلتر میانگین روی g اعمال می‌شود تا $\bar{g}$ حاصل شود. برای درک راحت‌تر این موضوع فرض می‌شود که هدف بخش‌بندی یک پایگاه داده به دو ناحیه باشد که ناحیه‌ی اول در مقادیر بالا قرار دارد. مجموعه‌ی T احتمال درست بودن تعلق داده‌ها به ناحیه‌ی اول را نشان می‌دهد. مطابق رابطه‌ی (۱)، هرچه اختلاف $\bar{g}(i,j)$ با $\bar{g}_{\min}$ بیشتر باشد $T(i,j)$ بیشتر خواهد بود. همچنین $F(i,j)$ و $I(i,j)$ به صورت زیر تعریف می-شوند [۱۰]:

$$F(i,j)=1-T(i,j) \qquad (۳)$$

$$I(i,j)=\frac{\delta(i,j)-\delta_{\min}}{\delta_{\max}-\delta_{\min}} \qquad (۴)$$

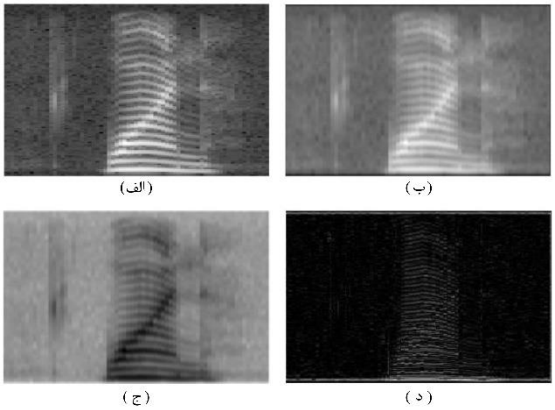
که در رابطه‌ی بالا $\delta_{\min}$ و $\delta_{\max}$ به ترتیب کمترین و بیشترین مقدار روی ماتریس δ بوده و $\delta(i,j)$ مطابق رابطه‌ی زیر تعریف می‌شود [۱۰]:

$$\delta(i,j)=\left|g(i,j)-\bar{g}(i,j)\right| \qquad (۵)$$

[20] Additive Margin MobileNet1D

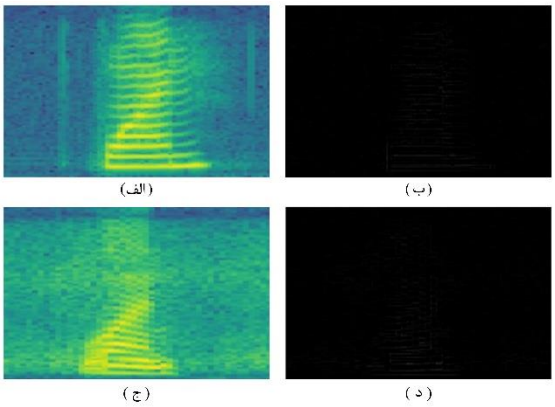
شکل (۲) نتیجه‌ی انتقال اسپکتروگرام به حوزه‌ی نتروسافیک کلاسیک را نشان می‌دهد. شکل (۲-الف) اسپکتروگرام یک سیگنال بدون نویز و شکل‌های (۲-ب)، (۲-ج) و (۲-د) به ترتیب مجموعه‌های T، F و I در حوزه‌ی نتروسافیک کلاسیک را نشان می‌دهند. همانطور که در شکل (۲-ب) مشاهده می‌شود مجموعه‌ی T، اسپکتروگرام را بلور می‌کند که اثر آن شبیه اثر یک فیلتر میانگین است. از طرفی مجموعه‌ی F عکس مجموعه‌ی T است. در مجموعه‌ی I هرجا اطلاعات فرکانسی مربوط به گوینده وجود دارد، مقدار بالایی ظاهر می شود (شکل ۲-د).

در انتقال به حوزه‌ی نتروسافیک، هدف مدل کردن نویز در اسپکتروگرام ورودی می‌باشد که در فضای نتروسافیک با مجموعه‌ی I نمایش داده می‌شود. در فضای نتروسافیک مجموعه‌ی عدم قطعیت I یکی از مهمترین مجموعه‌ها است که در پردازش سیگنال، معرف میزان نویز است.



شکل ۲- نتیجه‌ی انتقال اسپکتروگرام به حوزه‌ی نتروسافیک کلاسیک.

در روش‌های مرسوم محاسبه‌ی I در فضای نتروسافیک، هرچه اختلاف بین یک پیکسل با میانگین پیکسل‌های همسایه‌ی آن بیشتر باشد مقدار بزرگتری برای عدم قطعیت آن پیکسل در نظر گرفته می‌شود. با این تعریف، برای پیکسل‌های نزدیک به مرز در اسپکتروگرام، مقادیر بالایی به عدم قطعیت نسبت داده می‌شود در حالی که این پیکسل‌ها نویز نیستند. به عنوان مثال، شکل (۳-الف) تصویر یک اسپکتروگرام بدون نویز و شکل (۳-ج) تصویر اسپکتروگرام شکل (۳-الف) با اعمال نویز subway را نشان می‌دهد. شکل‌های (۳-ب) و (۳-د) به ترتیب مقدار عدم قطعیت اسپکتروگرام با استفاده از روش کلاسیک نتروسافیک، در حالت بدون نویز و نویز را نشان می‌دهند.



شکل ۳- انتقال اسپکتروگرام به فضای نتروسافیک کلاسیک، (الف) تصویر یک اسپکتروگرام بدون نویز، (ب) مقدار عدم قطعیت شکل الف، (ج) تصویر اسپکتروگرام شکل الف با اعمال نویز subway، (د) عدم قطعیت اسپکتروگرام شکل ج.

واضح است که روش کلاسیک نتروسافیک نمی‌تواند نویز موجود در اسپکتروگرام را به درستی مدل کند و بنابراین خروجی آن اطلاعات معنی‌داری نخواهد بود و به طور کلی اعمال روش کلاسیک نتروسافیک به تصاویر اسپکتروگرام به خروجی 0 منجر می‌شود. بر همین اساس اعمال آن به اسپکتروگرام شکل‌های (۳-الف) و (۳-ج)، شکل‌های (۳-ب) و (۳-د) را نتیجه می‌دهد که دو تصویر یکسان سیاه و بدون اطلاعات می‌باشند و می‌توان گفت که خروجی نتروسافیک برای دو تصویر بدون نویز و نویزی، تفاوت خاصی با هم ندارند. دلیل این امر این است که نویز اعمال شده در طول بعد فرکانسی اسپکتروگرام پخش می‌شود و به همین خاطر روش کلاسیک نتروسافیک که دارای یک فیلتر مربعی با ابعاد 3×3 است نمی‌تواند نویز موجود در اسپکتروگرام را مدل کند.

در اینجا به منظور حل این مشکل، تعریف جدیدی از مجموعه‌ی عدم قطعیت پیشنهاد می‌شود که در روابط زیر آمده است. به‌جای در نظر گرفتن اختلاف بین هر پیکسل و میانگین پیکسل‌های موجود در پنجره‌ی مربعی اطراف آن پیکسل، اختلاف هر پیکسل با پنجره مستطیلی با ابعاد a×b در نظر گرفته می‌شود که بُعد با اندازه‌ی a به راستای فرکانس تعلق دارد و $a=3b$. به این ترتیب برای مدل کردن نویز موجود در اسپکتروگرام، یک روش جدید نتروسافیک مطابق روابط زیر پیشنهاد می‌گردد:

$$T(i,j)=\frac{\bar{g}(i,j)}{\bar{g}_{mean}} \qquad (۶)$$

$$F(i,j)=1-T(i,j) \quad (۷)$$

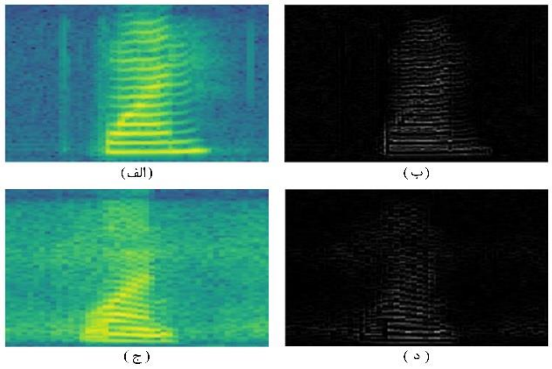
$$I(i,j)=\frac{\delta(i,j)}{\delta_{mean}} \quad (۸)$$

$$\bar{g}(i,j)=\frac{1}{ab}\sum_{m=-\frac{a}{2}}^{\frac{a}{2}}\sum_{n=-\frac{b}{2}}^{\frac{b}{2}} g(i+m,j+n) \quad (۹)$$

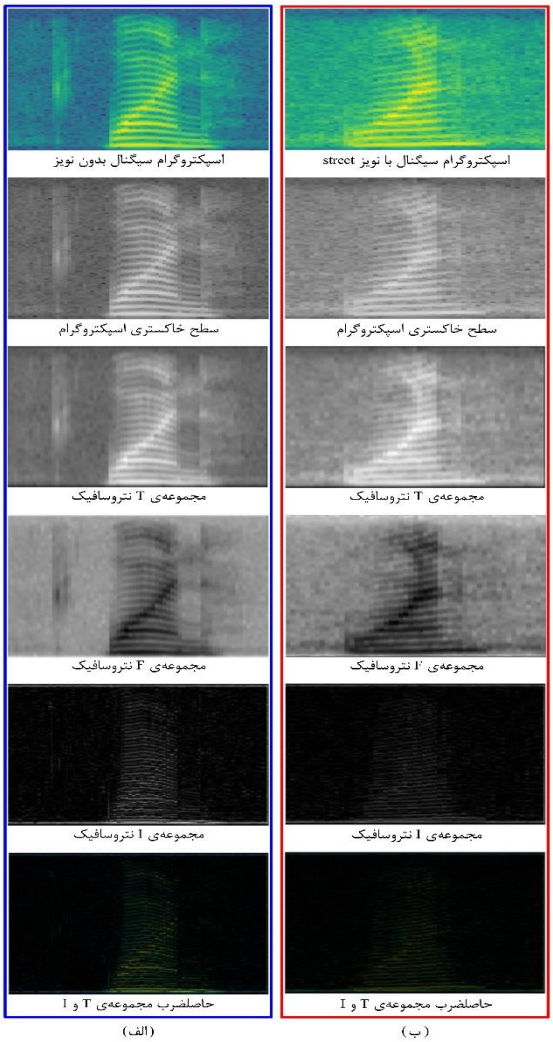
$$\delta(i,j)=\left|g(i,j)-\bar{g}(i,j)\right| \quad (۱۰)$$

واضح است که در رابطه‌ی (۹)، پنجره‌ی مستطیلی دارای ابعاد $a\times b$ می‌باشد. بر اساس روابط (۶) تا (۱۰)، عدم قطعیت اسپکتروگرام ورودی در شکل (۴) قابل مشاهده است. شکل‌های (۴-الف) و (۴-ج) به ترتیب تصویر دو اسپکتروگرام بدون نویز و با نویز مترو را نشان می‌دهند. شکل‌های (۴-ب) و (۴-د) به ترتیب عدم قطعیت اسپکتروگرام شکل‌های (۴-الف) و (۴-ج) را نشان می‌دهند. مشاهده می‌شود که روش پیشنهادی به خوبی توانسته است نویز را مدل کند.

شکل (۵) فرآیند اعمال روش پیشنهادی به اسپکتروگرام یک سیگنال بدون نویز و همان سیگنال با نویز street را نشان می‌دهد. تصاویر مربوط به سیگنال بدون نویز و سیگنال با نویز street به ترتیب در جعبه‌های آبی و قرمز قرار دارند. در مرحله‌ی اول تصویر اسپکتروگرام به سطح خاکستری تبدیل می‌شود سپس با انتقال اسپکتروگرام به فضای نتروسافیک، مجموعه‌های T، F و I نتروسافیک تشکیل می‌شود و در نهایت حاصلضرب مجموعه‌های T و I به عنوان اسپکتروگرام نهایی معرفی می‌شود. ابتدا سیگنال صوتی بدون نویز بررسی می‌شود؛ با توجه به شکل (۵-الف)، مجموعه‌ی T اسپکتروگرام را بلور می‌کند که اثر آن شبیه اثر یک فیلتر میانگین است. از طرفی، مجموعه‌ی F عکس مجموعه‌ی T است. همچنین بر اساس تصویر مجموعه‌ی I، هرجا اطلاعات فرکانسی مربوط به گوینده وجود دارد مقدار I در آنجا بالا خواهد بود. همانطور که در تصویر حاصلضرب دو مجموعه‌ی T و I دیده می‌شود مولفه‌های اصلی فرکانسی سیگنال صوتی در آن حفظ شده‌اند.



شکل ۴- انتقال اسپکتروگرام به فضای نتروسافیک با روش پیشنهادی، (الف) تصویر یک اسپکتروگرام بدون نویز، (ب) مقدار عدم قطعیت شکل الف، (ج) تصویر اسپکتروگرام شکل الف با اعمال نویز subway، (د) عدم قطعیت اسپکتروگرام شکل ج.



شکل ۵- فرآیند اعمال روش پیشنهادی به اسپکتروگرام یک سیگنال بدون نویز و همان سیگنال با نویز street، (الف) بدون نویز، (ب) با نویز.

اکنون اگر سیگنال (۵-الف) با نویز street آغشته شود مولفه‌های فرکانسی متعددی در سرتاسر اسپکتروگرام حاصل می‌شود که در شکل (۵-ب) به خوبی مشاهده می شود. در اینجا نشان داده می‌شود که چگونه این اسپکتروگرام به فضای نتروسافیک انتقال داده شده و نویز آن کنترل می‌گردد. مشابه حالت قبل، در مرحله‌ی اول اسپکتروگرام به سطح خاکستری تبدیل می‌شود، سپس مجموعه‌های نتروسافیک از آن استخراج می‌شوند.

همانطور که از شکل اسپکتروگرام مجموعه‌ی I مشخص است مقادیر این مجموعه برای فرکانس‌های نویزی نزدیک به صفر و برای نواحی غیرنویزی که اطلاعات اصلی سیگنال صوتی است مقادیر بالایی دارد که این از نکات کلیدی روش پیشنهادی است که بصورت اتوماتیک نویز را کنترل می‌کند. در نهایت در اسپکتروگرام نهایی (حاصلضرب مجموعه‌های T و I)، نشان داده شده است که در آن نواحی نویزی تا حد زیادی حذف شده‌اند و اطلاعات اصلی سیگنال باقی مانده است.

۳-۳-آنتروپی در حوزه‌ی نتروسافیک

در داده‌های در محدوده‌ی عددی خاصی، آنتروپی می‌تواند به عنوان معیاری برای بررسی توزیع مقادیر داده‌ها باشد. اگر مقادیر داده‌ها دارای توزیع غیریکنواخت باشد آنتروپی مینیمم است. آنتروپی در حوزه‌ی نتروسافیک از حاصل جمع آنتروپی مجموعه‌های T، F و I به صورت زیر بدست می‌آید [۷]:

$$En_{NS} = En_T + En_F + En_I \qquad (۱۱)$$

$$En_T = \sum_{i=\min(T)}^{\max(T)} p_T(i) \times \ln\left(p_T(i)\right) \qquad (۱۲)$$

$$En_F = \sum_{i=\min(F)}^{\max(F)} p_F(i) \times \ln\left(p_F(i)\right) \qquad (۱۳)$$

$$En_I = \sum_{i=\min(I)}^{\max(I)} p_I(i) \times \ln\left(p_I(i)\right) \qquad (۱۴)$$

که در روابط بالا $p_T(i)$ ، $p_F(i)$ و $p_I(i)$ احتمال عضویت عنصر i به مجموعه‌های نتروسافیک هستند.

۴-۳- عملگرهای میانگین آلفا و بهبود بتا

مقدار $I(i,j)$ برای اندازه‌گیری میزان عدم قطعیت عناصر به کار می‌رود. برای اینکه T و F با I مرتبط شود نیاز است در T و F تغییراتی ایجاد شود بطوریکه توزیع عناصر و آنتروپی I را تحت تاثیر قرار دهد. برای این کار میانگین آلفا تعریف می‌گردد. این عملگر بصورت زیر به تصویر اعمال می‌شود [۴۱]:

$$\bar{p}_{NS}(\alpha) = p(T_\alpha, F_\alpha, I_\alpha) \qquad (۱۵)$$

$$T_\alpha(i,j) = \begin{cases} T(i,j), & I(i,j) < \alpha \\ \bar{T}(i,j), & I(i,j) \geq \alpha \end{cases} \qquad (۱۶)$$

$$F_\alpha(i,j) = \begin{cases} F(i,j), & I(i,j) < \alpha \\ \bar{F}(i,j), & I(i,j) \geq \alpha \end{cases} \qquad (۱۷)$$

$$I_\alpha(i,j) = \frac{\bar{\delta}_T(i,j) - \bar{\delta}_{T\min}(i,j)}{\bar{\delta}_{T\max}(i,j) - \bar{\delta}_{T\min}(i,j)} \qquad (۱۸)$$

$$\bar{T}(i,j) = \frac{1}{w^2} \sum_{m=-\frac{w}{2}}^{\frac{w}{2}} \sum_{n=-\frac{w}{2}}^{\frac{w}{2}} T(i+m, j+n) \qquad (۱۹)$$

$$\bar{F}(i,j) = \frac{1}{w^2} \sum_{m=-\frac{w}{2}}^{\frac{w}{2}} \sum_{n=-\frac{w}{2}}^{\frac{w}{2}} F(i+m, j+n) \qquad (۲۰)$$

$$\bar{\delta}_T(i,j) = \left| \bar{T}(i,j) - \bar{\bar{T}}(i,j) \right| \qquad (۲۱)$$

$$\bar{\bar{T}}(i,j) = \frac{1}{w^2} \sum_{m=-\frac{w}{2}}^{\frac{w}{2}} \sum_{n=-\frac{w}{2}}^{\frac{w}{2}} \bar{T}(i+m, j+n) \qquad (۲۲)$$

بهتر است درجه‌ی عضویت داده‌ها در حد امکان از هم دور باشند. برای اینکار از عملگر بهبود بتا استفاده می‌شود که روابط آن در ادامه آمده است [۴۱]:

$$p'_{NS}(\beta) = p(T_\beta, F_\beta, I_\beta) \qquad (۲۳)$$

$$T_\beta(i,j) = \begin{cases} T(i,j), & I(i,j) < \alpha \\ T_\lambda(i,j), & I(i,j) \geq \alpha \end{cases} \qquad (۲۴)$$

$$F_\beta(i,j) = \begin{cases} F(i,j), & I(i,j) < \alpha \\ F_\lambda(i,j), & I(i,j) \geq \alpha \end{cases} \qquad (۲۵)$$

$$I_\beta(i,j) = \frac{\delta'_T(i,j) - \delta'_{T\min}(i,j)}{\delta'_{T\max}(i,j) - \delta'_{T\min}(i,j)} \qquad (۲۶)$$

$$T_\lambda(i,j) = \begin{cases} 2T^2(i,j), & I(i,j) < 0.5 \\ 1 - 2(1 - T(i,j))^2, & I(i,j) \geq \alpha \end{cases} \qquad (۲۷)$$

$$F_\lambda(i,j) = \begin{cases} 2F^2(i,j), & I(i,j) < 0.5 \\ 1 - 2(1 - F(i,j))^2, & I(i,j) \geq \alpha \end{cases} \qquad (۲۸)$$

$$\delta'_T(i,j)=\left|T_\lambda(i,j)-T'_\lambda(i,j)\right| \quad (۲۹)$$

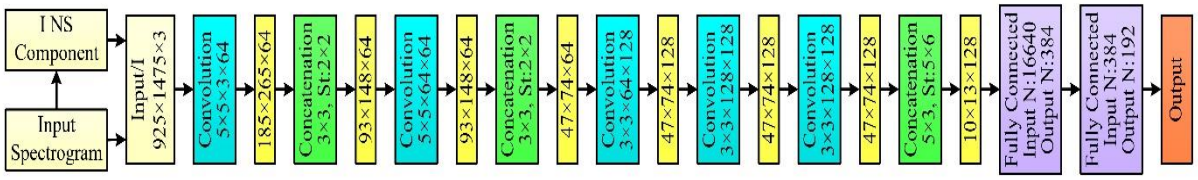
$$T'_\lambda(i,j)=\frac{1}{w^2}\sum_{m=-\frac{w}{2}}^{\frac{w}{2}}\sum_{n=-\frac{w}{2}}^{\frac{w}{2}}T_\lambda(i+m,j+n) \quad (۳۰)$$

۳-۵- شبکه‌ی عصبی کانولوشنال

اخیراً شبکه‌های عصبی کانولوشنال در بسیاری از کاربردها نتایج بسیار چشمگیری داشته است [۴۲،۴۳،۴۴،۴۵،۴۶] و به همین دلیل اینجا از شبکه‌های عصبی کانولوشن استفاده شده است. بعد از پیش‌پردازش داده‌ها و انتقال اسپکتروگرام به فضای نتروسافیک، سعی بر آن داریم که با تاثیر عدم قطعیت (مقدار داده‌ها در فضای نتروسافیک) در شبکه‌های عصبی کانولوشنال، عملکرد شبکه را بهبود دهیم. ابعاد داده ها در فضای نتروسافیک همانند ابعاد اصلی داده‌ها می باشد. به همین خاطر ابتدا با تاثیر عدم قطعیت داده‌ها روی داده‌های اصلی، مقدار نویز در ورودی را کاهش می دهیم. برای این منظور، مقدار ورودی بر مقدار عدم قطعیت تقسیم می‌شود. در حقیقت با این کار شبکه به گونه‌ای آموزش می‌بیند که داده‌هایی که عدم قطعیت پایینی دارند و در واقع نویز کمتری دارند در یک مقدار بیشتری ضرب شوند. نتیجه‌ی این کار این است که مقدار چنین داده‌هایی در شبکه تاثیر بیشتری خواهد داشت. به همین ترتیب داده‌هایی که عدم قطعیت بالایی دارند و به عبارتی احتمال نویزی بودن آنها بالا است در مقدار کمتری ضرب می‌شوند و در نتیجه تاثیر کمتری در شبکه خواهند داشت. شکل (۶) ساختار شبکه‌ی عصبی کانولوشنال پیشنهادی را نشان می‌دهد.

همانطور که در شکل (۶) مشاهده می‌شود لایه‌ی ورودی حاوی اسپکتروگرام صوت بوده و دارای ابعاد 3×1475×925 است. لایه‌ی اول، یک لایه‌ی کانولوشن بوده و دارای فیلترهایی با ابعاد 64×3×5×5 است. گام فیلترها جهت لغزش روی ورودی در بُعد فرکانس و زمان برابر 5 یعنی به صورت 5×5 بوده و بنابراین خروجی این لایه دارای ابعاد 64×295×185 خواهد بود. در شبکه‌های عصبی کانولوشنال معمولاً نمونه‌برداری‌های محلی از خروجی‌های لایه‌ی کانولوشن انجام می‌شود که این کار توسط لایه‌ی ادغام صورت می‌گیرد. بر این اساس، دومین لایه در این شبکه، لایه‌ی ادغام بوده و دارای ابعاد 3×3 با طول گام 2×2 است. این لایه، ابعاد را به نصف کاهش داده و ابعاد خروجی برابر 64×148×93 خواهد بود. سومین لایه، یک لایه‌ی کانولوشن بوده و فیلترهای آن دارای ابعاد 64×64×5×5 هستند. طول گام فیلترها در این لایه 1×1 است بنابراین این لایه، ابعاد ورودی را تغییر نداده و فقط اطلاعات محلی را به لایه‌ی بعدی انتقال می‌دهد. لایه‌ی چهارم از نوع ادغام بوده و برای کاهش بُعد استفاده می‌شود. همانند لایه‌ی دوم، این لایه دارای ابعاد 3×3 با طول گام 2×2 می‌باشد که بُعد ورودی را نصف می‌کند، بنابراین خروجی آن دارای ابعاد 64×74×47 است.

لایه‌های پنجم، ششم و هفتم از نوع کانولوشن بوده که فیلترهای آنها به ترتیب دارای ابعاد 128×64×3×3، 128×128×3×3 و 128×128×3×3 می‌باشند. لازم به ذکر است که طول گام در این لایه‌ها 1×1 است. با توجه به ابعاد فیلترها و طول گام، خروجی لایه‌های پنجم، ششم و هفتم با هم مساوی بوده و برابر 128×74×47 خواهد بود. لایه‌ی هشتم یک لایه‌ی ادغام با ابعاد 3×5 و طول گام 6×5 بوده و هدف این لایه کاهش بُعد می‌باشد. این لایه قبل از اولین لایه‌ی تماماً متصل قرار دارد و خروجی آن دارای ابعاد 128×13×10 است. لایه‌ی نهم اولین لایه‌ی تماماً متصل بوده که ورودی آن دارای 16640=128×13×10 و خروجی آن 384 نرون می‌باشد. لایه‌ی دهم آخرین لایه قبل از لایه‌ی خروجی بوده که ورودی و خروجی آن به ترتیب دارای 384 و 192 نرون می‌باشند. آخرین لایه، لایه‌ی خروجی و تعداد نرون‌های آن برابر تعداد کلاس‌های داده‌ها خواهد بود. به طور خلاصه، متغیرهای استفاده شده در این مقاله و تعریف آنها در جدول ۱ آمده است.



شکل ۶- ساختار شبکه‌ی عصبی کانولوشنال پیشنهادی (مستطیل‌های زردرنگ صرفاً ابعاد تصویر در آن مرحله از شبکه را نشان می‌دهند).

جدول ۱- متغیرهای استفاده شده در این مقاله و تعریف آنها.

متغیر	تعریف	متغیر	تعریف	متغیر	تعریف
$g(i,j)$	روشنایی پیکسل (i,j)	En_{NS}	آنتروپی در حوزه‌ی نتروسافیک	$\overline{F}$	میانگین F روی پنجره $w\times w$
P_{NS}	حوزه‌ی نتروسافیک	En_T	آنتروپی درست بودن	$\overline{\overline{T}}$	میانگین $\overline{T}$ روی پنجره $w\times w$
T	درصد درست بودن	En_F	آنتروپی نادرست بودن	$\overline{\delta}_T$	اختلاف $\overline{T}$ و $\overline{\overline{T}}$
I	درصد عدم قطعیت	En_I	آنتروپی عدم قطعیت	$p'_{NS}(\beta)$	عملگر بتا برای نتروسافیک
F	درصد نادرست بودن	$p_T(i)$	احتمال عضویت عنصر i به T	T_β	آنتروپی درست بودن بعد از اعمال عملگر بتا
$\overline{g}$	میانگین محلی g در پنجره $w\times w$	$p_F(i)$	احتمال عضویت عنصر i به F	F_β	آنتروپی نادرست بودن بعد از اعمال عملگر بتا
$\overline{g}_{\min}$	کمترین مقدار روی ماتریس $\overline{g}$	$p_I(i)$	احتمال عضویت عنصر i به مجموعه I	I_β	آنتروپی عدم قطعیت بعد از اعمال عملگر بتا
$\overline{g}_{\max}$	بیشترین مقدار روی ماتریس $\overline{g}$	$\overline{p}_{NS}(\alpha)$	عملگر میانگین آلفا برای مجموعه نتروسافیک	T_λ	تابع λ روی آنتروپی درست بودن
w	ابعاد پنجره	T_α	آنتروپی درست بودن بعد از اعمال عملگر آلفا	F_λ	تابع λ روی آنتروپی نادرست بودن
δ	قدرمطلق اختلاف g و $\overline{g}$	F_α	آنتروپی نادرست بودن بعد از اعمال عملگر آلفا	δ'_T	اختلاف T'_λ و T_λ
a	طول پنجره در روش پیشنهادی	I_α	آنتروپی عدم قطعیت بعد از اعمال عملگر آلفا	T'_λ	میانگین T_λ روی پنجره $w\times w$
b	عرض پنجره در روش پیشنهادی	$\overline{T}$	میانگین T روی پنجره با ابعاد $w\times w$		

۴- پایگاه‌های داده و تنظیم پارامترها

در این قسمت پایگاه‌های داده‌ی مورد استفاده در این مقاله تشریح می‌شوند. ابتدا پایگاه داده‌ی Aurora2 سپس پایگاه داده‌ی TIMIT بررسی شده و مشخصات آنها بیان خواهد شد.

۴-۱- پایگاه داده‌ی Aurora2

پایگاه داده‌ی Aurora2 حاوی گفتار متصل متشکل از ارقام انگلیسی با نرخ نمونه برداری 8kHz است [۴۷]. در این پایگاه داده دو مجموعه‌ی آموزشی و تست وجود دارد. داده‌های آموزش به دو دسته‌ی داده‌های تمیز و داده‌های نویزی تقسیم می‌شوند. داده‌های تست به سه دسته‌ی A، B و C تقسیم می‌شوند که هر کدام شامل گفتار تمیز و گفتار نویزی با نویزهای مختلف است. میزان سیگنال به نویز برای هر کدام از این مجموعه‌ها برابر 5، 15،10 و 20 دسی بل است.

مجموعه گفتار نویزی A دارای نویزهای مترو[۲۱]، ماشین[۲۲]، همهمه[۲۳] و نمایشگاه[۲۴] به همراه فیلتر کانال G.712 است. مجموعه گفتار نویزی B دارای نویزهای رستوران[۲۵]، خیابان[۲۶]، فرودگاه[۲۷] و ایستگاه قطار[۲۸] به همراه فیلتر کانال G.712 می‌باشد. مجموعه گفتار نویزی C دارای نویزهای مترو و خیابان است که هر کدام با نویزهای مجموعه‌ی A و B مشترک هستند. در مجموعه‌ی C از فیلتر کانال MIRS استفاده شده که متفاوت از نویز کانال مجموعه‌های A و B است.

۴-۲- پایگاه داده‌ی TIMIT

پایگاه داده‌ی TIMIT در بسیاری از سیستم‌های پردازش صوت به عنوان پایگاه داده‌ی مرجع برای ارزیابی استفاده شده است [۴۸]. این پایگاه داده شامل 630 گوینده از 8 منطقه‌ی آمریکا با لهجه‌های مختلف است که هر گوینده 10 جمله را ادا می‌کند بنابراین به طور کلی شامل 6300 جمله‌ی ادا شده توسط کل گویندگان می‌باشد. جدول ۲ مشخصات این پایگاه داده را بیان می‌کند.

[21] Subway
[22] Car
[23] Babble
[24] Exhibition
[25] Restaurant
[26] Street
[27] Airport
[28] Train Station

جدول ۲- توزیع تعداد گویندگان هر منطقه بر حسب جنس آنها.

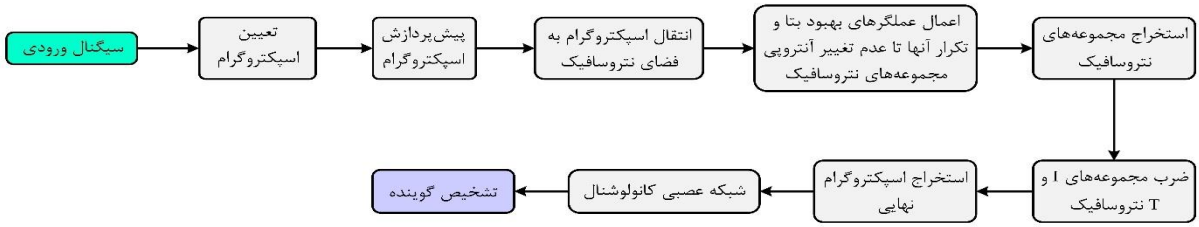
منطقه	مرد	زن	جمع
۱	۳۱ (۶۳٪)	۱۸ (۲۷٪)	۴۹ (۸٪)
۲	۷۱ (۷۰٪)	۳۱ (۳۰٪)	۱۰۲ (۱۶٪)
۳	۷۹ (۶۷٪)	۲۳ (۲۳٪)	۱۰۲ (۱۶٪)
۴	۶۹ (۶۹٪)	۳۱ (۳۱٪)	۱۰۰ (۱۶٪)
۵	۶۲ (۶۳٪)	۳۶ (۳۷٪)	۹۸ (۱۶٪)
۶	۳۰ (۶۵٪)	۱۶ (۳۵٪)	۴۶ (۷٪)
۷	۷۴ (۷۴٪)	۲۶ (۲۶٪)	۱۰۰ (۱۶٪)
۸	۲۲ (۶۷٪)	۱۱ (۳۳٪)	۳۳ (۵٪)
مجموع	۴۳۸ (۷۰٪)	۱۹۲ (۳۰٪)	۶۳۰ (۱۰۰٪)

جملاتی که توسط گویندگان ادا می‌شود با علائم SA، SX و SI برچسب زده می‌شوند. هر گوینده پنج جمله‌ی SX، سه جمله‌ی SI و دو جمله‌ی SA را ادا می‌کند. میانگین طول جملات 3 ثانیه بوده و همه‌ی جملات در محیط بدون نویز با فرکانس نمونه‌برداری 16kHz ضبط شده‌اند. این پایگاه داده شامل دو زیرمجموعه‌ی آموزش و تست است که معمولا بین 20% تا 30% داده‌ها به عنوان داده‌ی تست و 70% تا 80% داده‌ها به عنوان داده‌ی آموزش استفاده می شوند. تقسیم‌بندی پایگاه داده به مجموعه‌های آموزش و تست باید به گونه‌ای باشد که هیچ گوینده‌ای در هر دو زیرمجموعه‌ی آموزش و تست قرار نگیرد و باید از تمام 8 ناحیه در هر دو زیرمجموعه‌ی آموزش و تست استفاده شود.

۴-۳- مدل سیستم و مفروضات

با توجه به مراحل ذکر شده در بخش ۳، مدل سیستم پیشنهادی در شکل (۷) آمده است.

مفروضات مورد استفاده در مدل پیشنهادی به شرح زیر می‌باشد: در شبکه‌های عصبی کانولوشنال، پارامترهای مربوط به وزن‌ها به صورت تصادفی مقداردهی اولیه می شوند. در فرایند آموزش، از داده‌های تمیز و نویزی به صورت جداگانه استفاده می‌گردد. همچنین توابع فعال خروجی خطی بوده و در آموزش شبکه، از گرادیان نزولی تصادفی[29] و اندازه‌ی دسته‌های کوچک[30] استفاده می‌شود. به دلیل وجود مقادیر مثبت و منفی در ورودی و خروجی، در لایه‌ی کانولوشن تابع فعال تانژانت هایپربولیک استفاده شده است. تعداد گام‌های[31] مورد نیاز برای آموزش شبکه 100 می‌باشد که بعد از 100 تکرار، دقت شبکه پایین می‌آید و overfitting رخ می‌دهد. اندازه‌ی دسته[32] به 128 تنظیم شده و تعداد تکرار[33] بر اساس اندازه‌ی داده‌های آموزشی تعیین می‌گردد. لازم به ذکر است که نرخ یادگیری شبکه، 0.0001 در نظر گرفته شده است. همچنین ابعاد پنجره‌ی مستطیلی برابر $a=30, b=10$ می‌باشد.



شکل ۷- مدل سیستم پیشنهادی.

معروف‌ترین روش برای تقسیم‌بندی داده به آموزش و تست، روش ارزیابی k-fold نام دارد. در این روش داده‌ها به k دسته‌ی مساوی تقسیم می‌شوند و در k مرحله سیستم آموزش می‌بیند و تست می‌شود. در هر مرحله، k-1 دسته برای آموزش و 1 دسته‌ی باقیمانده برای تست استفاده می‌گردد. 10-flold از روش‌های مرجع برای ارزیابی می‌باشد که سیستم در 10 مرحله آموزش می‌بیند و تست می‌شود. در اینجا از همین روش یعنی 10-fold استفاده شده است.

۵- نتایج پیاده‌سازی

به منظور ارزیابی روش پیشنهادی، از دو پایگاه داده‌ی مذکور استفاده شده است. نتایج دقت روش پیشنهادی و روش‌های [۳۰]، [۵]، [۲۵] و [۵۰] روی پایگاه داده‌ی

[29] Stochastic Gradient Descent (SGD)
[30] Mini Batch
[31] Epoch
[32] Batch Size
[33] Iteration

Aurora2 در جدول ۳ و شکل (۸) آمده است.

بر اساس نتایج جدول ۳ و شکل (۸) برای پایگاه داده‌ی Aurora2 می‌توان نتیجه گرفت: ۱) روش پیشنهادی نسبت به روش‌های [۳۰]، [۵]، [۲۵] و [۵۰]، از 30 حالت مختلف در 23 حالت (76.67% موارد) دارای دقت بالاتری بوده است؛ ۲) متوسط دقت روش پیشنهادی روی همه‌ی حالت-ها 93.79% است که در مقایسه با روش‌های [۳۰]، [۵]، [۲۵] و [۵۰]، به ترتیب 15.59%، 11.03%، 4.66% و 2.05 دقت بالاتری دارد؛ ۳) در بین حالت‌های مختلف، بیشترین دقت متعلق به روش پیشنهادی بوده و برابر 98.53% است؛ ۴) در بین روش‌های قبلی، روش [۵۰] بهتر از بقیه بوده است؛ ۵) دقت روش پیشنهادی روی همه‌ی حالت‌های بخش C بالاتر از سایر روش‌ها است. از طرفی در این بخش، دقت روش پیشنهادی با سایر روش‌ها فاصله‌ی زیادی دارد.

برای پایگاه داده‌ی TIMIT، نتایج دقت روش پیشنهادی و روش‌های [۳۰]، [۵]، [۲۵] و [۵۰]، برای سیگنال تمیز و نویزی در جدول ۴ و شکل (۹) آمده است. نویز با SNRهای 0 dB، 5 dB، 10 dB و 15 dB به سیگنال تمیز اضافه می‌شود. ذکر این نکته ضروری است که سیگنال نویز مربوط به پایگاه داده‌ی NOISEX-92 [۴۹] است.

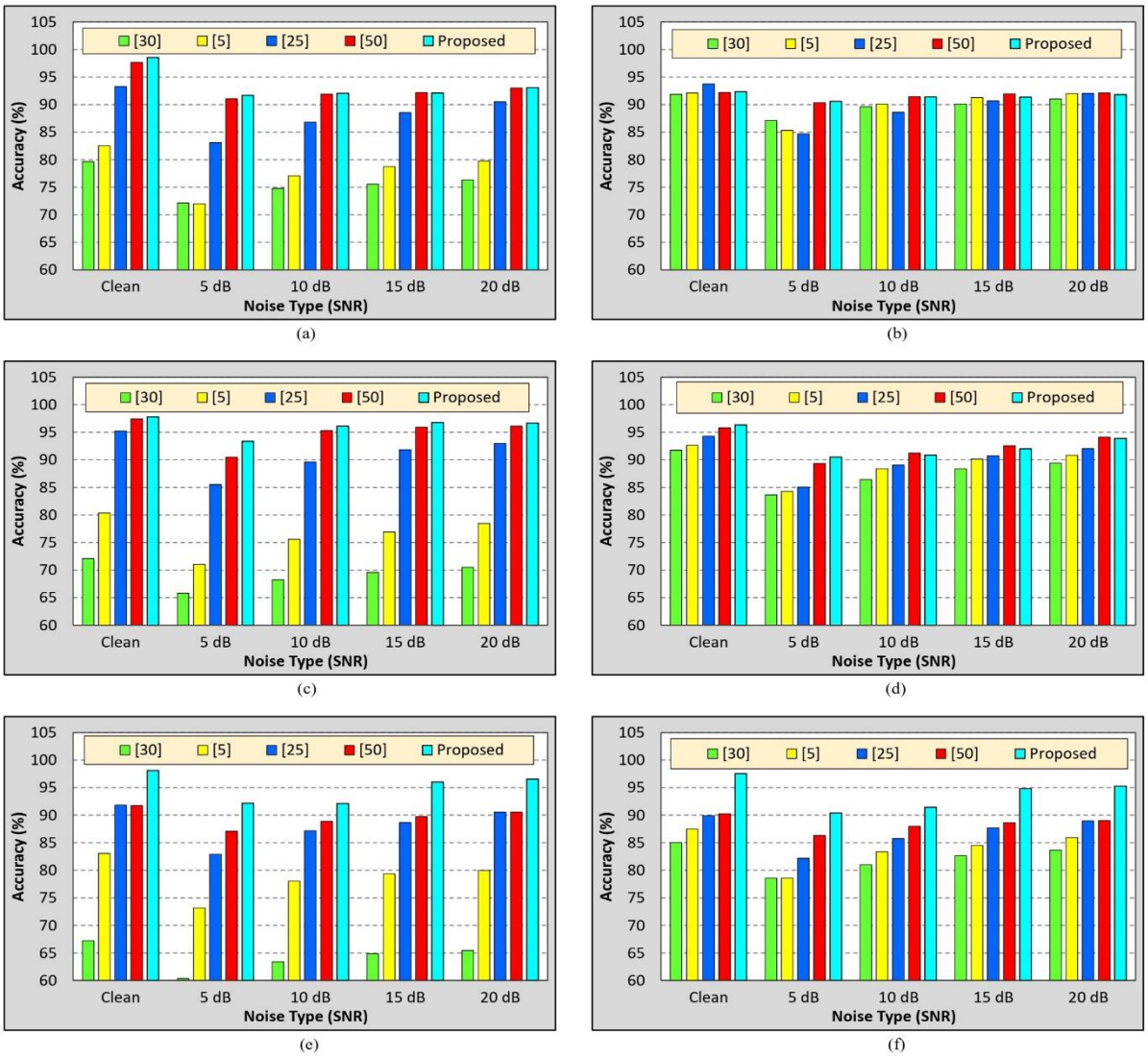
برای پایگاه داده‌ی TIMIT، می‌توان نتایج جدول ۴ و شکل (۹) را به این صورت جمع‌بندی کرد: ۱) روش پیشنهادی نسبت به روش‌های [۳۰]، [۵]، [۲۵] و [۵۰]، در 44 حالت از 61 حالت متفاوت (72.13% موارد) دقت بالاتری کسب کرده است؛ ۲) روش [۵۰] برای 3 حالت از 4 حالت نویز Leopard دقت بالاتری داشته و بهتر از روش پیشنهادی بوده است؛ ۳) متوسط دقت روش پیشنهادی برای همه‌ی حالت‌ها 95.24% می‌باشد که در مقایسه با روش‌های [۳۰]، [۵]، [۲۵] و [۵۰]، به ترتیب 23.65%، 13.77%، 2.61% و 1.27% بالاتر بوده است؛ ۴) بعد از روش پیشنهادی، روش [۵۰] بهتر از سایر روش‌ها عمل نموده است. در حقیقت در 10 حالت از 61 حالت (16.39% موارد) دقت بالاتری داشته است.

جدول ۳- نتایج دقت روش پیشنهادی روی پایگاه داده‌ی Aurora2.

Dataset Section	Training Data	Noise Type (SNR)	[30]	[5]	[25]	[50]	Proposed
A	Clean	Clean	79.61	82.52	93.28	97.66	**98.53**
		5 dB	72.15	72.00	83.09	91.05	**91.66**
		10 dB	74.79	77.07	86.81	91.88	**92.03**
		15 dB	75.55	78.73	88.53	**92.16**	**92.11**
		20 dB	76.34	79.76	90.49	92.99	**93.06**
A	Noisy	Clean	91.86	92.11	**93.71**	92.19	92.33
		5 dB	87.11	85.29	84.68	90.32	**90.58**
		10 dB	89.58	90.07	88.60	**91.44**	91.39
		15 dB	90.09	91.26	90.67	**91.94**	91.36
		20 dB	91.01	91.98	92.05	**92.09**	91.78
B	Clean	Clean	72.07	80.37	95.22	97.42	**97.81**
		5 dB	65.79	71.03	85.51	90.48	**93.38**
		10 dB	68.22	75.58	89.62	95.29	**96.13**
		15 dB	69.60	76.93	91.86	95.94	**96.78**
		20 dB	70.48	78.46	93.01	96.12	**96.69**
B	Noisy	Clean	91.75	92.64	94.31	95.81	**96.35**
		5 dB	83.66	84.27	85.04	89.32	**90.52**
		10 dB	86.41	88.41	89.05	**91.23**	90.84
		15 dB	88.38	90.14	90.73	**92.56**	91.99
		20 dB	89.39	90.82	92.02	**94.12**	93.91
C	Clean	Clean	67.21	83.04	91.83	91.76	**98.12**
		5 dB	60.35	73.17	82.91	87.09	**92.19**
		10 dB	63.40	78.03	87.18	88.89	**92.12**
		15 dB	64.85	79.36	88.67	89.73	**96.03**
		20 dB	65.44	79.96	90.58	90.57	**96.57**
C	Noisy	Clean	85.00	87.49	89.91	90.26	**97.55**
		5 dB	78.57	78.60	82.20	86.31	**90.40**
		10 dB	81.02	83.37	85.77	88.00	**91.43**
		15 dB	82.64	84.52	87.69	88.63	**94.83**
		20 dB	83.65	85.95	88.94	89.01	**95.28**
Average	-	-	78.20	82.76	89.13	91.74	**93.79**

جدول ۴- نتایج دقت روش پیشنهادی روی پایگاه داده‌ی TIMIT.

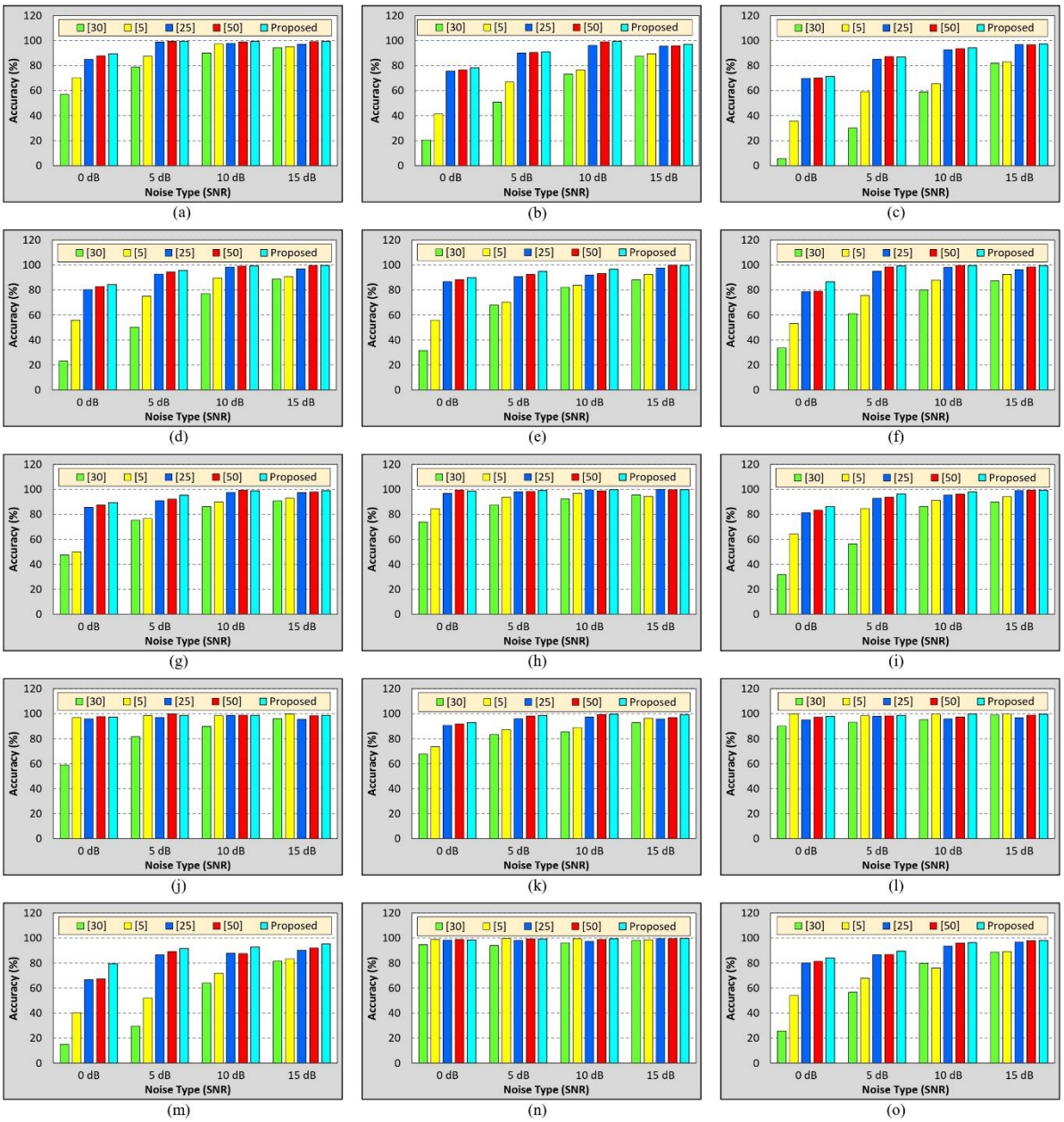
Noise Type	Noise level (SNR)	[30]	[5]	[25]	[50]	Proposed
Clean	-	97.1	97.64	97.25	**98.84**	98.71
Babble	0 dB	56.93	70.45	85.03	87.92	**89.57**
	5 dB	78.76	87.65	98.88	99.42	**99.61**
	10 dB	90.13	97.39	97.93	99.01	**99.35**
	15 dB	94.22	95.1	97.2	99.29	**99.33**
Buccaneer1	0 dB	20.51	41.33	75.7	76.68	**78.29**
	5 dB	50.83	66.99	90.06	90.55	91.08
	10 dB	73.23	76.5	96.23	99.21	**99.34**
	15 dB	87.52	89.61	95.67	95.95	**97.12**
Buccaneer2	0 dB	5.68	35.67	69.91	70.07	**71.25**
	5 dB	30.16	59.18	85.16	**87.31**	86.94
	10 dB	58.93	65.55	92.78	93.63	**94.28**
	15 dB	81.92	82.98	96.91	96.78	**97.42**
Dest.Eng.	0 dB	23.25	55.84	80.32	82.82	**84.38**
	5 dB	50.29	75.12	92.81	94.5	**95.73**
	10 dB	77.12	89.56	98.27	99.05	**99.26**
	15 dB	88.75	90.76	97.07	**99.63**	99.56
Dest. Ops	0 dB	31.52	55.91	86.63	88.33	**89.82**
	5 dB	67.99	70.33	90.63	92.4	**94.79**
	10 dB	82.24	83.81	91.81	93.32	**96.64**
	15 dB	88.09	92.46	97.61	**99.86**	99.71
F16	0 dB	33.7	53.25	78.59	79.03	**86.52**
	5 dB	61.03	75.49	95.11	98.43	**99.25**
	10 dB	80.06	87.84	98.16	**99.61**	99.57
	15 dB	87.3	92.54	96.38	98.41	**99.47**
Factory1	0 dB	47.5	50.13	85.6	87.35	**89.46**
	5 dB	75.3	76.64	90.95	92.16	**95.27**
	10 dB	86.15	89.89	97.7	**99.27**	98.88
	15 dB	90.78	93	97.46	97.88	**98.94**
Factory2	0 dB	73.84	84.55	96.73	**99.32**	98.57
	5 dB	87.7	93.74	98.14	98.34	**99.04**
	10 dB	92.38	96.78	99.54	98.81	**99.68**
	15 dB	95.56	94.42	**99.88**	99.62	99.75
Hfchannel	0 dB	31.81	64.32	81.3	83.39	**86.33**
	5 dB	56.33	84.66	92.85	93.75	**96.42**
	10 dB	86.27	91.23	95.43	96.09	**98.15**
	15 dB	89.7	94.16	99.01	99.12	**99.31**
Leopard	0 dB	58.91	97	95.82	**97.69**	97.28
	5 dB	81.51	98.55	97.02	**99.81**	98.61
	10 dB	89.86	98.33	98.64	**98.88**	98.70
	15 dB	95.99	**99.88**	95.55	98.35	98.65
M109	0 dB	67.73	73.49	90.76	91.81	**92.84**
	5 dB	83.33	87.17	95.9	98.2	**98.47**
	10 dB	85.45	88.91	97.27	99.23	**99.58**
	15 dB	92.81	96.34	95.75	96.88	**99.35**
Machinegun	0 dB	90.2	**99.98**	95.09	97.17	97.81
	5 dB	93.16	98.35	97.87	98.29	**98.65**
	10 dB	95.13	99.51	95.81	97.43	**99.73**
	15 dB	98.95	**99.57**	96.74	98.87	99.53
Volvo	0 dB	14.96	40.13	66.64	67.42	**79.44**
	5 dB	29.49	52.04	86.49	89.2	**91.67**
	10 dB	63.96	71.76	87.93	87.55	**92.83**
	15 dB	81.52	83.23	90.19	92.09	**95.34**
Pink	0 dB	94.7	**98.78**	98.21	98.76	98.43
	5 dB	93.99	**99.82**	98.08	99.29	99.21
	10 dB	96.2	99.37	97.46	98.76	**99.41**
	15 dB	98.09	98.64	99.5	99.46	**99.67**
White	0 dB	25.67	53.95	80.15	81.4	**84.12**
	5 dB	56.52	67.74	86.64	86.88	**89.53**
	10 dB	79.73	75.98	93.57	95.96	**96.21**
	15 dB	88.72	89.16	96.83	97.88	**98.02**
Average	-	71.59	81.48	92.63	93.97	**95.24**



شکل ۸- نتایج دقت روی پایگاه داده‌ی Aurora2 برای حالت‌های مختلف، a) بخش A پایگاه داده و داده‌های Clean برای آموزش، b) بخش A پایگاه داده و داده‌های Noisy برای آموزش، c) بخش B پایگاه داده و داده‌های Clean برای آموزش، d) بخش B پایگاه داده و داده‌های Noisy برای آموزش، e) بخش C پایگاه داده و داده‌های Clean برای آموزش، f) بخش C پایگاه داده و داده‌های Noisy برای آموزش.

۶- محدودیت‌های مدل پیشنهادی و مسیر آینده‌ی تحقیق

روش ارائه شده برای انتقال اسپکتروگرام سیگنال‌های صوتی به فضای نتروسافیک به گونه‌ای طراحی شده است که نویزهای موجود در اسپکتروگرام را به خوبی آشکار می کند. هدف اصلی این است که بتوان اثر این نویزها را در مراحل بعدی کم کرد. با توجه به اینکه نویزهای متنوعی از قبیل خیابان، فرودگاه، رستوران، ایستگاه قطار، ماشین، همهمه و مترو در اسپکتروگرام سیگنال‌های صوتی وجود دارد، محدودیت اصلی روش پیشنهادی این است که نمی تواند همه‌ی این نویزها را در فضای نتروسافیک آشکار کند. بنابراین، مسیر آینده این تحقیق به این صورت پیشنهاد می‌شود که برای هر نوع نویز، یک روش انتقال مجزا در فضای نتروسافیک ارائه شود. برای این کار می‌توان نویزهای مصنوعی به اسپکتروگرام سیگنال‌های صوتی تمیز تزریق کرد و سپس عملگرهای انتقال در فضای نتروسافیک را بصورت اختصاصی برای شناسایی این نویزها طراحی نمود. با انجام این کار، اسپکتروگرام‌های مختلفی از یک سیگنال صوتی حاصل می‌شود که در نهایت می‌توان با تجمیع یا اشتراک‌گیری از آنها، به یک اسپکتروگرام واحد رسید که حاوی داده‌های دقیق‌تری از مولفه‌های فرکانسی یک سیگنال صوتی باشد.



شکل ۹- نتایج دقت روی پایگاه داده‌ی TIMIT برای نویزهای مختلف؛ a) Babble، b) Buccaneer1، c) Buccaneer2، d) Destroyer Engine، e) Destroyer Ops، f) F16، g) Factory 1، h) Factory 2، i) Hfchannel، j) Leopard، k) M109، l) Machinegun، m) Volvo، n) Pink، o) White.

۷- نتیجه‌گیری

در این مقاله، یک سیستم تشخیص گوینده مبتنی بر نتروسافیک و شبکه‌های عصبی کانولوشنال پیشنهاد شد. در روش پیشنهادی، ابتدا اسپکتروگرام سیگنال صوتی تشکیل شد، سپس اسپکتروگرام سیگنال به فضای نتروسافیک انتقال داده شد. در مرحله‌ی بعد عملگرهای بهبود بتا به مجموعه‌های نتروسافیک اعمال گردید و این عملیات تا ثابت شدن آنتروپی مجموعه‌های نتروسافیک تکرار می‌شد. در مرحله‌ی آخر فرآیند تشخیص گوینده توسط یک شبکه‌ی عصبی کانولوشنال به پایان می‌رسید. برای ارزیابی و بررسی کارایی روش پیشنهادی، دو پایگاه داده‌ی معروف Aurora2 و TIMIT استفاده شد. نتایج مربوط به دقت تشخیص گوینده نشان داد روش پیشنهادی قابل رقابت با روش‌های اخیر است.

مراجع

[1] Müller, C., Speaker Classification I: Fundamentals, Features, and Methods, Springer-Verlag Berlin Heidelberg, 2007.

[2] Ajmera, P. K., and Holambe, R. S., "Fractional Fourier transform based features for speaker recognition using support vector machine," Computers & Electrical Engineering, Vol. 39, No. 2, pp. 550-557, 2013.

[3] Rathor, S., and Jadon, R. S., "Text indpendent speaker recognition using wavelet cepstral coefficient and butter worth filter," In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5, 2017.

[4] Lei, H., Meyer, B. T., and Mirghafori, N., "Spectro-temporal Gabor features for speaker recognition," In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4241-4244, 2012.

[5] Qi, M., Yu, Y., Tang, Y., Deng, Q., Mai, F., and Zhaxi, N., "Deep CNN with se block for speaker recognition," In 2020 Information Communication Technologies Conference (ICTC), pp. 240-244, 2020.

[6] Zhang, M., Zhang, L., and Cheng, H. D., "A neutrosophic approach to image segmentation based on watershed method," Signal Processing, Vol. 90, No. 5, pp. 1510-1517, 2010.

[7] Heshmati, A., Gholami, M., and Rashno, A., "Scheme for unsupervised colour–texture image segmentation using neutrosophic set and non-subsampled contourlet transform," IET Image Processing, Vol. 10, No. 6, pp. 464-473, 2016.

[8] Guo, Y., Şengür, A., and Ye, J., "A novel image thresholding algorithm based on neutrosophic similarity score," Measurement, Vol. 58, pp. 175-186, 2014.

[9] Guo, Y., and Şengür, A., "A novel image edge detection algorithm based on neutrosophic set," Computers & Electrical Engineering, Vol. 40, No. 8, pp. 3-25, 2014.

[10] Y. Guo and A. Sengur, "Ncm: Neutrosophic c-means clustering algorithm," Pattern Recognition, vol. 48, no. 8, pp. 2710–2724, 2015.

[11] Ohi, A. Q., Mridha, M. F., Hamid, M. A., and Monowar, M. M., "Deep speaker recognition: Process, progress, and challenges," IEEE Access, Vol. 9, pp. 89619-89643, 2021.

[12] Kabir, M. M., Mridha, M. F., Shin, J., Jahan, I., and Ohi, A. Q., "A survey of speaker recognition: Fundamental theories, recognition methods and opportunities," IEEE Access, Vol. 9, pp. 79236-79263, 2021.

[13] Keshet, J., and Bengio, S. (Eds.), Automatic speech and speaker recognition: Large margin and kernel methods, John Wiley & Sons, 2009.

[14] Hanifa, R. M., Isa, K., and Mohamad, S., "A review on speaker recognition: Technology and challenges," Computers & Electrical Engineering, Vol. 90, 107005, 2021.

[15] Bai, Z., and Zhang, X. L., "Speaker recognition based on deep learning: An overview," Neural Networks, Vol. 140, pp. 65-99, 2021.

[16] Bharath, K. P., and Kumar, R., "Multitaper based MFCC feature extraction for robust speaker recognition system," In 2019 Innovations in Power and Advanced Computing Technologies (i-PACT), Vol. 1, pp. 1-5, 2018.

[17] Ghalamiosgouei, S., and Geravanchizadeh, M., "Robust Speaker Identification Based on Binaural Masks," Speech Communication, Vol. 132, pp. 1-9, 2021.

[18] Liu, Z., Wu, Z., Li, T., Li, J., and Shen, C., "GMM and CNN hybrid method for short utterance speaker recognition," IEEE Transactions on Industrial informatics, Vol. 14, No. 7, pp. 3244-3252, 2018.

[19] Han, J. H., Bae, K. M., Hong, S. K., Park, H., Kwak, J. H., Wang, H. S., ... and Lee, K. J., "Machine learning-based self-powered acoustic sensor for speaker recognition," Nano Energy, Vol. 53, pp. 658-665, 2018.

[20] Sahidullah, M., and Saha, G., "A novel windowing technique for efficient computation of MFCC for speaker recognition," IEEE signal processing letters, Vol. 20, No. 2, pp. 149-152, 2012.

[21] Chowdhury, A., and Ross, A., "Fusing MFCC and LPC features using 1D triplet CNN for speaker recognition in severely degraded audio signals," IEEE transactions on information forensics and security, Vol. 15, pp. 1616-1629, 2019.

[22] Devi, K. J., Singh, N. H., and Thongam, K., "Automatic speaker recognition from speech signals using self organizing feature map and hybrid neural network," Microprocessors and Microsystems, Vol. 79, 103264, 2020.

[23] Jahangir, R., Teh, Y. W., Memon, N. A., Mujtaba, G., Zareei, M., Ishtiaq, U., ... and Ali, I., “Text-independent speaker identification through feature fusion and deep neural network,” IEEE Access, Vol. 8, pp. 32187-32202, 2020.

[24] Nunes, J. A. C., Macêdo, D., and Zanchettin, C., “Additive margin sincnet for speaker recognition,” In 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1-5, 2019.

[25] Moumin, A. A., and Kumar, S. S., “Automatic Speaker Recognition using Deep Neural Network Classifiers,” In 2021 2nd International Conference on Computation, Automation and Knowledge Management (ICCAKM), pp. 282-286, 2021.

[26] Chien, J. T., and Peng, K. T., “Neural adversarial learning for speaker recognition,” Computer Speech & Language, Vol. 58, pp. 422-440, 2019.

[27] Zhang, X., Zou, X., Sun, M., Zheng, T. F., Jia, C., and Wang, Y., “Noise robust speaker recognition based on adaptive frame weighting in GMM for i-vector extraction,” IEEE Access, Vol. 7, pp. 27874-27882, 2019.

[28] Dai, M., Dai, G., Wu, Y., Xia, Y., Shen, F., and Zhang, H., “An Improved Feature Fusion for Speaker Recognition,” In 2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC), pp. 183-187, 2019.

[29] Schädler, Marc René, Bernd T. Meyer, and Birger Kollmeier. "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition." The Journal of the Acoustical Society of America, Vol. 131, No. 5, pp. 4134-4151, 2012.

[30] Rashno, E., Akbari, A., and Nasersharif, B., “A convolutional neural network model based on neutrosophy for noisy speech recognition,” In 2019 4th International Conference on Pattern Recognition and Image Analysis (IPRIA), pp. 87-92, 2019.

[31] Bahmaninezhad, F., Zhang, C., and Hansen, J. H., “An investigation of domain adaptation in speaker embedding space for speaker recognition,” Speech Communication, Vol. 129, pp. 7-16, 2021.

[32] Mesgarani, N., Slaney, M., and Shamma, S. A., “Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations,” IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 3, pp. 920-930, 2006.

[33] Ahmed, A. I., Chiverton, J. P., Ndzi, D. L., and Becerra, V. M., “Speaker recognition using PCA-based feature transformation,” Speech Communication, Vol. 110, pp. 33-46, 2019.

[34] Xu, J., Li, S., Jiang, J., and Dou, Y., “A simplified speaker recognition system based on FPGA platform,” IEEE Access, Vol. 8, pp. 1507-1516, 2019.

[35] Chakroun, R., and Frikha, M., “Robust text-independent speaker recognition with short utterances using Gaussian mixture models,” In 2020 International Wireless Communications and Mobile Computing (IWCMC), pp. 2204-2209, 2020.

[36] Bian, T., Chen, F., and Xu, L., “Self-attention based speaker recognition using Cluster-Range Loss,” Neurocomputing, Vol. 368, pp. 59-68, 2019.

[37] Avila, A. R., O’Shaughnessy, D., and Falk, T. H., “Automatic speaker verification from affective speech using Gaussian mixture model based estimation of neutral speech characteristics,” Speech Communication, Vol. 132, pp. 21-31, 2021.

[38] Lin, T., and Zhang, Y., “Speaker recognition based on long-term acoustic features with analysis sparse representation,” IEEE Access, Vol. 7, pp. 87439-87447, 2019.

[39] Nunes, J. A. C., Macêdo, D., and Zanchettin, C., “Am-mobilenet1d: A portable model for speaker recognition,” In 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1-8, 2020.

[40] Govindan, S. M., Duraisamy, P., and Yuan, X., “Adaptive wavelet shrinkage for noise robust speaker recognition,” Digital Signal Processing, Vol. 33, pp. 180-190, 2014.

[41] Guo, Yanhui, and Heng-Da Cheng., “New neutrosophic approach to image segmentation,” Pattern Recognition, Vol. 42, No. 5, pp. 587-595, 2009.

[۴۲] مرضیه زارع نظری؛ محسن سرداری زارچی؛ سیما عمادی؛ هادی پورمحمدی، "چارچوبی برای استخراج آناتومی و طبقه بندی تصاویر پشه با رویکرد یادگیری عمیق"، مدل‌سازی در مهندسی، دوره ۲۰، شماره ۷۰، مهر ۱۴۰۱، صفحه ۱۲۰-۱۰۷.

[۴۳] میثم عفتی، رحمت مدندوست، و زینب فلاح زرجو بازکیایی، "ارزیابی عملکرد مدل های شبکه عصبی مصنوعی، نروفازی و رگرسیون چند متغیره در پیش بینی مقاومت فشاری بتن به کمک روش بارنقطه ای"، مدل‌سازی در مهندسی، دوره ۱۸، شماره ۶۲، پاییز ۱۳۹۹، صفحه ۱۱۳-۹۹.

[۴۴] محمدحسین ولایتی، "ارزیابی قابلیت ضریب مشارکت ژنراتورها به منظور تعیین نوع نوسانات سیگنال کوچک سیستم قدرت با استفاده از روش‌های تحلیلی و پیش‌بینی همزمان آن‌ها با استفاده از شبکه عصبی"، مدل‌سازی در مهندسی، دوره ۱۳، شماره ۴۲، پاییز ۱۳۹۴، صفحه ۱۳۳-۱۱۹.

[۴۵] مسلم سردشتی بیرجندی؛ حسین رحمانی؛ سعید فراهت، "کاربرد شبکه های عصبی عمیق در طبقه بندی تصاویر آسیب های شبکه فاضلاب و مشخص کردن مسیرهای بحرانی آنها"، مدل‌سازی در مهندسی، دوره ۲۰، شماره ۷۰، مهر ۱۴۰۱، صفحه ۱۳۲-۱۲۱.

[۴۶] سیاوش حسینی، سعید ستایشی، غلامحسین روشنی، عبدالحمید زاهدی و فرزین شماع، "افزایش کارآیی جریان سنج دوفازی با استفاده از روش های استخراج ویژگی حوزه ی فرکانس و شبکه عصبی در طیف خروجی آشکارساز"، مدل‌سازی در مهندسی، دوره ۱۹، شماره ۶۷، زمستان ۱۴۰۰.

[47] Hirsch, H. G., and Pearce, D., “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” In ASR2000-Automatic speech recognition: challenges for the new Millenium ISCA tutorial and research workshop (ITRW), 2000.

[48] TIMIT dataset, available online on: https://catalog.ldc.upenn.edu/LDC93S1. Last accessed at 14 September 2021.

[49] NOISEX-92 noise dataset, available online on: http://spib.linse.ufsc.br/noise.html. Last accessed at 14 September 2021.

[۵۰] عبدالرضا رشنو؛ صادق فدایی؛ عبدالصمد حمیدی، "تشخیص خودکار گوینده مبتنی بر ویژگی های استخراج شده از بانک فیلتر گابور و شبکه های عصبی کانولوشنال"، *مدل‌سازی در مهندسی، شناسه دیجیتالی:* 10.22075/JME.2022.26690.2245.