Journal of Acoustical Engineering Society of Iran, Vol. 10, No. 2, 2023

(Research Article)
Telephone robustness speaker verification using time delay neural network

M. Asgari *, N. Akbari, M. Aghagolzade, M. S. Mehrabikia
Broadcast Engineering Department, Islamic Republic of Iran Broadcasting University (IRIBU)

Received: 2021/12/15, Accepted: 2022/12/28

Abstract

In this research, TDNN model and x-vector are presented in order to robust noise and frequency filtering caused by telephone communication. MFCC is used as the speaker-related audio feature as input to this model. The output of neural network of this model is considered as an x-vector so that it can be used in the decision stage. In the decision stage, PLDA was used for scoring and comparison. In order to increase accuracy and reduce EER, the training dataset is a combination of relatively clean VoxCeleb 1,2 dataset and Callhome telephone dataset, as well as noise and telephone dataset obtained from the data augmentation method. The results of using this method for EER in the clean state are 3.09%, which has improved about 0.15% (3.24% has been obtained in previous works) in the worst case and 6.93% (10.2% has been obtained in previous works) in the best case compared to the base models. When training with Voxceleb1,2 and Callhome datasets was used as an adaptation, the EER was 4.95%. In the worst case, when only the Voxceleb1 data is converted to a telephone, the EER is 14.34%.

Keywords: Speaker verification, Time delay neural network, x-Vector, Mel frequency cepstral coefficients, Probability linear discriminant analysis.

pp. 11-20 (In Persian)

* Corresponding author E-mail: m.asgari@iribu.ac.ir

مقاله پژوهشی

مجله انجمن مهندسی صوتیّات ایران
سال ۱۴۰۱، دوره: ۱۰، شماره: ۲، صفحات: ۱۱-۲۰

تأیید هویت گویندهٔ مقاوم به شرایط تلفنی با استفاده از شبکهٔ عصبی تأخیر زمانی

محمد عسگری*، نرگس اکبری، مهران آقاگلزاده، محمدصادق محرابی‌کیا
دانشکده فنی و مهندسی رسانه، دانشگاه صدا و سیمای جمهوری اسلامی ایران
دریافت: ۱۴۰۰/۰۹/۲۴، پذیرش: ۱۴۰۱/۱۰/۰۷

چکیده

در این تحقیق الگوی شبکهٔ عصبی تأخیر زمانی و بردار-x به منظور مقاوم‌سازی در برابر نوفه و صافی کردن (فیلترینگ) بسامدی ناشی از ارتباط تلفنی ارایه گردیده است. از ضرایب کپسترال بسامد مل به عنوان ویژگی صوتی مرتبط با گوینده به عنوان ورودی این الگو استفاده شده است. خروجی شبکهٔ عصبی این الگو به‌صورت یک بردار-x در نظر گرفته شده است تا بتواند در مرحلهٔ تصمیم‌گیری از آن استفاده شود. در مرحلهٔ تصمیم‌گیری از تحلیل تفکیک خطی احتمالاتی به‌منظور امتیازدهی و مقایسه استفاده شده است. به‌منظور افزایش دقت و کاهش نرخ خطای برابر، داده‌های آموزشی ترکیبی از دادگان نسبتاً تمیز وُکس‌سِلِب۱، ۲ و دادگان تلفنی کال‌هوم و هم‌چنین دادگان نوفه‌ای و تلفنی به‌دست آمده از روش داده‌افزایی هستند. نتایج حاصل از به‌کارگیری این شیوه برای نرخ خطای برابر در حالت تمیز ۳٫۰۹ درصد است که نسبت به الگوهای پایه در بدترین حالت در حدود ۰٫۱۵ درصد (۳٫۲۴ درصد در کارهای پیشین به‌دست آمده است) و در بهترین حالت ۶٫۹۳ درصد (۱۰٫۲ درصد در کارهای پیشین به‌دست آمده است) بهبود یافته است. در زمانی که آموزش با دادگان وُکس‌سِلِب۱، ۲ و دادگان کال‌هوم به عنوان تطبیق استفاده شده است، نرخ خطای برابر معادل ۴٫۹۵ درصد به‌دست آمده است. در بدترین حالت زمانی که فقط دادگان وُکس‌سِلِب۱ به تلفنی تبدیل شده، نرخ خطای برابر معادل با ۱۴٫۳۴ درصد شده است.

کلیدواژه‌ها: تأیید هویت گوینده، شبکهٔ عصبی تأخیر زمانی، بردار-x، ضرایب کپسترال بسامد مل، احتمال تحلیل تفکیک خطی.

۱. مقدمه

در این تحقیق، هدف بررسی تأیید هویت یک گوینده در حالت تمیز با حالت تلفنی است. نمونه‌های گفتاری یک مکالمهٔ تلفنی از یک فرد، با صدای او در حالت تمیز در زمان ضبط به‌صورت میکروفونی مقایسه می‌شوند. در این شرایط، تشخیص گوینده که صدای او از تلفن عبور کرده با حالت صدای تمیز او باعث تشخیص اشتباه فرد شنونده مسئول مقایسهٔ این نمونه‌ها یا سامانه تأیید هویت گوینده می‌شود. تشخیص هویت گوینده به دو شکل کاملاً خودکار و نیمه- خودکار[1] انجام می‌گیرد. تحقیقات متعددی در مورد سامانه‌های تایید گوینده تا به حال صورت گرفته که حوزه‌های کاربردی مختلف نظیر بانک‌داری، تایید هویت در کاربردهای جرم‌شناسی قانونی و سامانه‌های امنیتی را شامل می‌شود [۱-۲]. عملکرد این سامانه‌ها با توجه به قرار گرفتن گوینده در شرایط مختلف محیطی مانند محیط‌های نوفه‌ای (نوفه خیابان، ماشین، فرودگاه، مترو، قطار، همهمه[2] و غیره)، شرایط گویشی مانند گفتار با احساس نسبت به حالت طبیعی و گفتار با لهجه‌های مختلف (نظیر تهرانی، کردی، لری، آذری و غیره) که در آن صدای گوینده به‌صورت تمیز ضبط شده است، کاهش می‌یابد [۳-۴]. اغلب تحقیقات در حوزهٔ تأیید هویت گوینده در شرایط کاملاً تمیز صورت گرفته است در صورتی‌که بسیاری از فعالیت‌ها مانند کاربردهای مجرمانه از طریق تهدید تلفنی همراه با نوفه‌های محیطی پس‌زمینه است. نسخه جدید شبکهٔ عصبی عمیق (دی‌اِن‌اِن)[3] با استفاده از بردار- x، ابزار منعطف و در عین حال

* نویسنده پاسخگو: m.asgari@iribu.ac.ir

[1] Semi-automatic
[2] Babel
[3] DNN; Deep Neural Network

قدرتمندی برای تشخیص خودکار گوینده فراهم می‌کند و باعث می‌شود که متخصصان حوزه جرم‌شناسی تفسیر بهتری از نتایج شناسایی با استفاده از چارچوب کاری نرخ شباهت (اِل‌آر)[1] ارائه نمایند. برای این منظور، معیار درست‌نمایی [۵] معرفی شد، که متخصصان برای بیان استحکام یافته‌های خود از آن استفاده می‌کنند [۶-۷]. بنابراین، فعالیت متخصص صدا، تخمین احتمال تفاوت مشاهده‌شده از اندازه‌گیری بین نمونهٔ گفتارها با فرض این‌که آن‌ها توسط همان گوینده و یا گوینده‌های مختلف صحبت شده [۸]، می‌باشد. هدف اصلی این تحقیق بررسی عملکرد تأیید هویت گوینده در شرایط گفتار تلفنی است. طبق جدول ۱ خلاصه‌ای از ورودی‌های متداول و شبکه‌های عصبی برای استخراج ویژگی گوینده مبتنی بر جاساز عمیق ارائه شده است که از آن می‌توان دریافت که شبکه‌های عصبی مبتنی بر شبکهٔ عصبی پیچشی (سی‌اِن‌اِن)[2] و ویژگی‌های صوتی بانک صافی (فیلتر)[3]/ ضرایب کپسترال بسامد مل (اِم‌اِف‌سی‌سی)[4] محبوبیت دارند، در حالی که برخی از ساختارهای پیچشی دو-بعدی، به عنوان مثال شبکهٔ عصبی باقی‌مانده، از اسپکتروگرام[5] به عنوان ویژگی ورودی استفاده می‌کنند. علاوه بر ورودی‌های متداول فوق، مانند ضرایب کپسترال بسامد مل، طیف و بانک صافی مل، لیو و همکاران [۹] اخیراً ارزیابی مجدد گسترده‌ای از استخراج‌کنندهٔ چهارده ویژگی صوتی ارائه داده‌اند. آن‌ها دریافتند که ویژگی‌های صوتی مجهز به فن‌های مرکزهای طیفی[6]، توابع تأخیر گروهی و مهار نوفه تجمعی گزینه‌های امیدوارکننده‌ای را برای ضرایب کپسترال بسامد مل فراهم می‌کنند. از این‌رو ساختار شبکه نقشی اساسی در عملکرد دارد. شبکهٔ عصبی تأخیر زمانی قابل‌تعمیم[7] و شبکهٔ عصبی تأخیر زمانی عامل‌گیری‌شده (فاکتورگیری‌شده)[8] به طور قابل توجهی نرخ خطای برابر را در مجموعه داده اِس‌آی‌تی‌دابلیو[9] کاهش داده‌اند [۱۰]، که شبکهٔ عصبی تأخیر زمانی عامل‌گیری‌شده بیش از ۴۰ درصد کاهش نسبی نرخ خطای برابر نسبت به شبکهٔ عصبی تأخیر زمانی اصلی را به‌دست می‌آورد. این امر بیانگر اهمیت ساختار شبکه در عملکرد است [۱۱-۱۲]. با توجه به این‌که هدف از این تحقیق، ارائه روشی برای تأیید هویت گوینده در شرایط تلفنی است، پس علاوه بر ارائه روشی که با دادهٔ کم بتواند هویت گوینده را تشخیص دهد، باید با شرایط تلفنی نیز مقابله کرد. بیش‌تر کاربردهای تأیید هویت گوینده در شرایط تلفنی و نوفه‌ای است در صورتی‌که بیش‌تر تحقیقات انجام‌شده در شرایط نسبتاً تمیز صورت گرفته است. پس در این تحقیق، معماری تأیید هویت گویندهٔ مقاوم به شرایط تلفنی با استفاده از شبکهٔ عصبی تأخیر زمانی بررسی شده است که شامل جزئیات دادگان وُکس‌سِلِب۱، ۲[10]، کال‌هوم[11]، پیش‌پردازش، استخراج ویژگی، الگو، بخش تصمیم‌گیری، داده‌افزایی و غیره است. هم‌چنین در پیش‌پردازش از آشکارساز فعالیت صوتی (وی‌اِی‌دی)[12] استفاده شده است. سپس ویژگی‌های صوتی با استفاده از ضرایب کپسترال بسامد مل استخراج شده و به عنوان ورودی به شبکهٔ عصبی تأخیر زمانی ارائه می‌شود. خروجی شبکهٔ عصبی بردار- x است که به بخش تصمیم‌گیری داده می‌شود. در بخش تصمیم‌گیری، با توجه به امتیاز به‌دست‌آمده و آستانهٔ از پیش تعیین‌شده، تصمیم گرفته می‌شود که آیا گویندهٔ مورد نظر با گویندهٔ ثبت-نام‌شده از قبل، یکسان است یا خیر. معیارهای ارزیابی نظیر نرخ خطای برابر، کمینه تابع هزینه تصمیم (دی‌اِف‌سی)[13]، نمودارهای توزیع فراوانی احتمال، پراکندگی امتیازات و منحنی مشخصه عملکرد برای بیان نتیجه در این تحقیق به‌کار گرفته شده است.

۲. مواد و روش‌ها

در این تحقیق عملیات تطبیق امواج صوتی، با روی‌کرد تأیید هویت گوینده با روی‌کرد مستقل از متن پیاده‌سازی گردیده است. بلوک‌های اصلی روش پیشنهادی شامل: (الف) واحد تجزیه‌وتحلیل- واحد استخراج شبه‌سنج‌های

[1] LR; Likelihood Rate
[2] CNN; Convolutional Neural Network
[3] Filter bank
[4] MFCC; Mel Frequency Cepstral Coefficients
[5] Spectrogram
[6] Spectral centroids
[7] Extended time delay neural network
[8] Factorized time delay neural network
[9] SITW

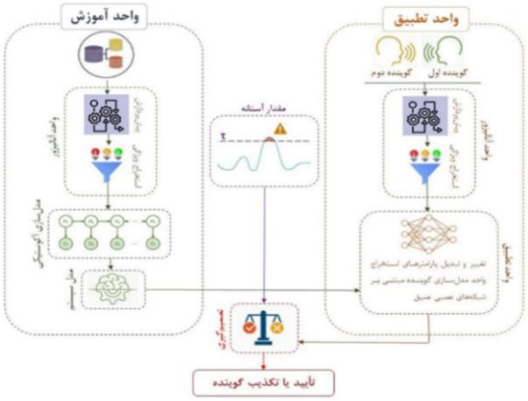
[10] VoxCeleb1,2
[11] Callhome
[12] VAD; Voice Activity Detection
[13] DCF; Decision Cost Function

جدول ۱ خلاصه‌ای از ورودی‌ها و ساختارهای شبکه در استخراج ویژگی گوینده عمیق [۱۲].

ورودی‌ها	شبکهٔ عصبی پیچشی	شبکهٔ عصبی کوتاه مدت ماندگار (اِل‌اِس‌تی‌اِم)[1]	ساختارهای ترکیبی[2]
شکل موج خالص	دیگر	----------	سی‌اِن‌اِن- اِل‌اِس‌تی‌اِم سی‌اِن‌اِن- آرجی‌یو[3]
اسپکتروگرام	رِس‌نِت[4] وی‌جی‌جی‌نِت[5] اینسپِشِن- رِسنِت- وی۱[6]	----------	سی‌اِن‌اِن- آرجی‌یو
طیف در نوارهای بسامدی مختلف	تی‌دی‌اِن‌اِن رِس‌نِت وی‌جی‌جی‌نِت اینسپِشِن- رِسنِت- وی۱ دیگر	[۱۲]	بی‌اِل‌اِس‌تی‌اِم[7]- رِس‌نِت تی‌دی‌اِن‌اِن- اِل‌اِس‌تی‌اِم
ضرایب کپسترال بسامد مل	تی‌دی‌اِن‌اِن رِس‌نِت دیگر	----------	تی‌دی‌اِن‌اِن- اِل‌اِس‌تی‌اِم

[1] LSTM; Long-Short Term Memory
[2] Hybrid structures
[3] GRU
[4] ResNet
[5] VGGNet
[6] Inception-resnet-v1
[7] BLSTM

علامت گفتار و (ب) واحد تطبیق- تغییر و تبدیل شبه‌سنج‌های استخراج واحد الگوسازی گوینده مبتنی بر شبکهٔ عصبی تأخیر زمانی (تی‌دی‌اِن‌اِن)[1] است که در شکل ۱ نشان داده شده است. در ادامه هر بلوک به‌صورت کامل شرح داده خواهد شد.



شکل ۱ روش پیشنهادی به‌منظور تشخیص هویت گویندهٔ مستقل از متن.

[1] TDNN; Time Delay Neural Network

۲-۱. دادگان

وُکس‌سِلِب [۱۳] یکی از دادگان پرکاربرد با تعداد گویندگانی بیش از ۷ هزار نفر و ۱ میلیون قطعه گفتاری برای کاربردهای تأیید هویت گوینده است که در بیش‌تر تحقیقات این حوزه مورد استفاده قرار می‌گیرد. هم‌چنین دادگان مذکور از ویدئوهای جمع‌آوری‌شده از یوتیوب[2] (شامل مصاحبه‌های بازیگران و هنرمندان) تهیه شده که دارای نوفه‌های محیطی، صداهای مختلف پس‌زمینه و غیره با ضبط میکروفونی است. پس با توجه به وسعت گوینده، گفتار و شرایط محیطی که این مجموعه دارد، در این تحقیق نیز از این دادگان استفاده شده است. هم‌چنین به منظور ایجاد شرایط تلفنی، از دادگان کال‌هوم [۱۴] به منظور تطبیق استفاده شده است. در جهت داده‌افزایی، دادگان وُکس‌سِلِب۱ با کاهش پهنای نوار و سه مرتبه فیلترینگ به‌طور مستقل، به‌منظور تلفنی کردن به‌کار گرفته شده و از آن در تعدادی از آزمایش‌ها برای آموزش استفاده شده است. به‌طور کلی تعداد دادگان به‌کار

[2] Youtube

جدول ۲ مشخصات دادگان ترکیب‌شده در این تحقیق به منظور تأیید هویت گوینده در شرایط نوفه‌ای-تلفنی.

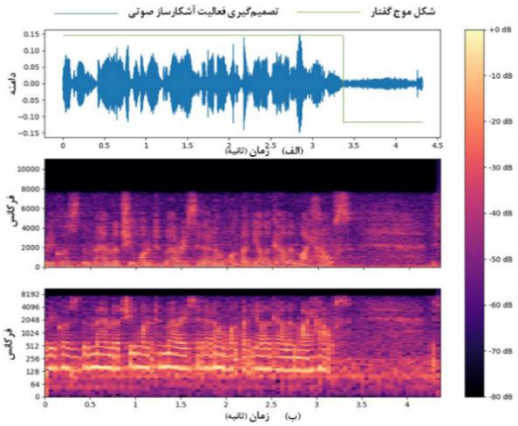
نام دادگان	نوع داده	نوع دسترسی	تعداد گویندگان	تعداد قطعات گفتاری
[۱۳] *Vox۱+Vox۲	میکروفونی/ نوفه‌ای	رایگان	۷۳۶۳	۱۱۰۰۰۰۰
کال‌هوم [۱۴]	تلفنی	رایگان	۱۲۰	۸۰۰۰
مجموع	-	-	۷۴۸۳	۱۱۰۸۰۰۰

* منظور از Vox۱+ Vox۲ ترکیب دادگان وُکس‌سِلِب۱ و وُکس‌سِلِب۲ به‌منظور آموزش است.

برده‌شده در این تحقیق برای آموزش و تطبیق به تفکیک در جدول ۲ نشان داده شده است.

۲-۲. پیش‌پردازش

انسان‌ها در تمایز بین صداهای گفتاری و غیرگفتاری خوب عمل می‌نمایند، که این نیز بخشی اساسی از تشخیص گوینده در کاربردهای مختلف نظیر خانه‌های هوشمند، تشخیص جرم و غیره از طریق شنیدن می‌باشد. در سامانه‌های خودکار نیز مطلوب است به‌منظور کاهش حجم محاسبات و صرفه‌جویی در زمان پردازش، ویژگی‌ها فقط از بخش‌های گفتاری شکل موج صوتی استخراج شوند، که این امر نیاز به آشکارساز فعالیت صوتی [۱۵-۱۶] دارد. وقتی بخش‌های صوتی بسیار نوفه‌ای و یا تخریب‌شده هستند، بخش‌های گفتاری بسیار مهم می‌شوند.



شکل ۲ (الف) شکل موج گفتار با تصمیم‌گیرهای فعالیت گفتار (مقادیر یک و صفر نشان‌دهنده گفتار و سکوت می‌باشند)، (ب) اسپکتروگرام شکل موج مربوطه.

عملکرد آشکارساز فعالیت صوتی در شکل ۲-الف نشان داده شده است، جایی که حضور و یا عدم‌حضور گفتار توسط یک علامت باینری پوشانده‌شده روی نمونه‌های گفتار نشان داده شده است. طیف‌سنجی گفتار مربوطه در شکل ۲-ب نشان داده شده است. در این کار از پنجرهٔ پاوی[1] مطابق با منبع [۱۵] استفاده شده است (فرمول ۱). پنجرهٔ پاوی در لبه‌ها به آرامی صفر می‌شود که از پس‌دهی[2] اجزای بسامد پایین با انرژی بالا به قسمت‌های بسامد بالای طیف جلوگیری می‌کند.

$$g(n) = ((۱ - \cos(n / N۲\pi)) / ۲)^{۰/۸۵} \qquad [۱۵] \quad (۱)$$

۲-۳. استخراج ویژگی‌ها

این ویژگی‌ها به شبه‌سنج‌های استخراج‌شده از بخش‌های کوتاه و قاب‌های (فریم‌های) گفتار با طول زمانی ۲۰ الی ۲۵ میلی‌ثانیه اشاره دارد. محبوب‌ترین ویژگی‌های امواج صوتی کوتاه‌مدت عبارتند از: ضرایب کپسترال بسامد مل (فرمول‌های ۲ الی ۵) [۱۷] و ویژگی‌های کدگذاری مبتنی بر پیش بینی خطی (اِل‌پی‌سی)[3] [۱۸]. در این‌جا به‌طور خلاصه در مورد ویژگی‌های ضرایب کپسترال بسامد مل توضیح داده خواهد شد. برای به‌دست آوردن این ضرایب از گفتار، ابتدا نمونه‌های صوتی به بخش‌های هم‌پوشانی کوتاه‌مدت تقسیم می‌شوند. یک قاب علامت گفتار ۲۵ میلی‌ثانیه معمولی در شکل ۳-الف نشان داده شده است. علامت به‌دست‌آمده در این بخش‌ها و قاب‌ها پس از آن در پنجره ضرب می‌شوند تا طیف توان فوریه به‌دست آید. در مرحلهٔ بعد، لگاریتم طیف محاسبه شده و سپس بانک صافی (فیلتر) با پهنا و فاصلهٔ غیرخطی مورد تحلیل واقع می‌گردد. لگاریتم مقیاس ضرایب را گسترش می‌دهد و هم‌چنین عمل ضرب را به جمع تقلیل می‌دهد.

تجزیه و تحلیل بانک صافی، انرژی طیف را در هر نوار بسامدی، هم‌چنین به‌عنوان ضرایب انرژی بانک صافی شناخته می‌شود، نشان می‌دهد. یک بانک صافی ۲۴

[1] Povey window
[2] Bleeding
[3] LPC; Linear Predictive Coding

مجرای معمولی و خروجی‌های آن در شکل ۳-ج و ۳-د به‌ترتیب نمایش داده شده‌اند. همان‌طور که در این‌جا مشهود است، بانک صافی به گونه‌ای طراحی شده است که نسبت به تغییرات بسامد در انتهای پایین طیف، مشابه سامانه شنوایی انسان حساس‌تر باشد [۱۷]. سرانجام، ضرایب کپسترال بسامد مل‌ها با انجام یک تبدیل کسینوسی گسسته[1] روی شبه‌سنج‌های انرژی بانک صافی و حفظ تعدادی از ضرایب اولیه به‌دست می‌آیند. به طور کلی، شبه‌سنج‌های سرعت و شتاب محاسبه‌شده در چندین قاب گفتار به ضرایب کپسترال بسامد مل اضافه می‌شوند. این شبه‌سنج‌ها که به ترتیب به‌عنوان دلتا و دلتا دلتا شناخته می‌شوند، نشان‌دهندهٔ خواص پویایی ضرایب ویژگی کوتاه‌مدت هستند. فنون الگوسازی که با هدف یادگیری رفتار تخریب‌ها از بیان گفتاری بهره می‌برند در بهبود استحکام مؤثرتر می‌باشند. به‌عنوان مثال، یک سامانه خودکار که چندین نمونه صدا از گوینده‌های مختلف را در سروصدای کنار جاده ضبط کرده است، در تشخیص گوینده‌ها در آن محیط بهتر خواهد بود. پس در این تحقیق نیز با اضافه کردن شرایط تلفنی، به عنوان دادگان آموزش و تطبیق، سعی بر شبیه‌سازی شرایط، به‌منظور بهبود دقت تأیید هویت گوینده شده است.

$$f_{Mel} = ۲۵۹۵\log_{۱۰}\left(۱+\left(f/۷۰۰\right)\right) \quad (۲)$$

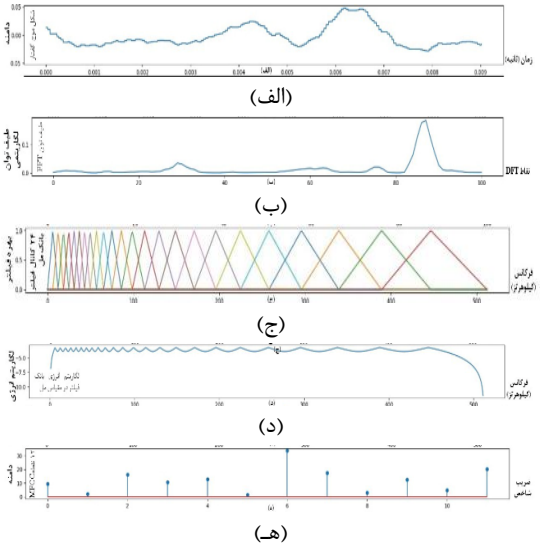
$$s(m) = \sum_{۰}^{N-۱}\left[\left|X(k)\right|^{۲} H_m(k)\right]; ۰ \le m \le M-۱ \quad (۳)$$

$$s(m) = \sum_{۰}^{N-۱}\left[\left|X(k)\right|^{۲} H_m(k)\right]; \quad ۰ \le m \le M-۱ \quad H_m(k) = \begin{cases} ۰, & k < f(m-۱) \\ \dfrac{۲(k-f(m-۱))}{f(m)-f(m-۱)}, & f(m-۱) < k \le f(m) \\ \dfrac{۲(f(m+۱)-k)}{f(m+۱)-f(m)}, & f(m) < k \le f(m+۱) \\ ۰, & k > f(m+۱) \end{cases} \quad (۴)$$

$$c(n) = \sum_{m=۰}^{M-۱}\log_{۱۰}\left(s(m)\right)\cos\left(\frac{\pi n(m-۰/۵)}{M}\right); n = ۰,۱,۲,\ldots,C-۱ \quad (۵)$$

در روابط (۲) تا (۵) داریم، f بسامد فیزیکی بر مبنای هرتز، M تعداد کل صافی‌های وزنی مل مثلثی، $H_m(k)$ وزن گرفته‌شده از k-اُمین بین[2] اسپکترام[3] انرژی به m-اُمین نوار خروجی (صافی بانک مل) [۱۷].

[1] DCT; Discrete Cosine Transform
[2] Bin
[3] Spectrum



شکل ۳ مراحل استخراج شاخص ضرایب کپسترال بسامد مل از روی قالب‌های گفتار؛ (الف) قالب ۲۰۰ نمونه‌ای نمایشگر گفتار ۲۵ میلی‌ثانیه‌ای نمونه‌برداری‌شده با نرخ ۸ کیلوهرتز، (ب) چگالی طیف توان که ۱۰۱ نقطه اول را نشان می‌دهد، (ج) بانک صافی مل مثلثی با ۲۴ مجرا، (د) انرژی خروجی بانک صافی لگاریتمی از بانک صافی مل و (هـ) ۱۲ ضریب ایستان ضرایب کپسترال بسامد مل که از روی تبدیل کسینوسی گسسته گرفته‌شده از ضرایب انرژی بانک صافی گرفته‌شده و در ۱۲ مقدار اولیه خود را نگه داشته است.

۲-۴. بردار-x

بردار دیگر ویژگی مستخرج از لایه مخفی، بردار-x نامیده می‌شود [۱۹]. این ویژگی براساس درون جاگرفته‌های[4] شبکهٔ عصبی عمیق است که از معماری چندلایه شبکهٔ عصبی تأخیر زمانی (با لایه‌های کاملاً متصل) و با بافت موقتی[5] در هر لایه (که قاب نامیده می‌شوند) بهره برده است. معماری درونی شبکهٔ عصبی تأخیر زمانی در جدول ۳ آورده شده است. پنج لایه اول با بافت موقتی کوچکی که در مرکز قاب جاری یعنی t قرار گرفته است، روی قاب‌های گفتار عمل می‌کند. برای مثال به‌خاطر بافت موقتی لایه‌های اولیه، قاب ۳ فقط ۱۵ قاب می‌بیند. پس از آموزش شناسه‌های گوینده به‌عنوان بردارهای هدف، خروجی لایه قطعه[6] شمارهٔ ۶ (بردار-x) به‌عنوان ورودی دسته‌بند تحلیل تفکیک خطی احتمالاتی (پی‌اِل‌دی‌اِی)[7] به‌کار گرفته شده است. برای گفتارهای[8] کوتاه در منبع

[4] Embeddings
[5] Temporal context
[6] Segment
[7] PLDA; Probalistic Linear Discrete Analysis
[8] Utterance

جدول ۳ تعداد و مشخصات لایه‌های روش پیشنهادی در این تحقیق.

لایه	متن لایه	متن کلی	ورودی	خروجی
فریم ۱	[t-۲،t+۲]	۵	۱۲۰	۵۱۲
فریم ۲	{t-۲،t،t+۲}	۹	۱۵۳۶	۵۱۲
فریم ۳	{t-۳،t،t+۳}	۱۵	۱۵۳۶	۵۱۲
فریم ۴	{t}	۱۵	۵۱۲	۵۱۲
فریم ۵	{t}	۱۵	۵۱۲	۱۵۰۰
تجمیع	[۰،T)	T	۱۵۰۰T	۳۰۰۰
بخش‌بندی ۶	{۰}	T	۳۰۰۰	۵۱۲
بخش‌بندی ۷	{۰}	T	۵۱۲	۵۱۲
خروجی (سافت‌مکس*)	{۰}	T	۵۱۲	N

* Softmax

[۲۰] آمده است که با تغییر ابعاد لایه‌های ششم و هفتم تا ۱۵۰، می‌تواند مدت‌زمان کوتاه‌تری از قاب تطابق یابد، اما با وجود ابعاد کم‌تر، مقدار نرخ خطای برابر بالاتری روی گفتارهای طولانی‌تر اولیه ایجاد کرده است.

۲-۵. واحد تصمیم‌گیری

به‌منظور تصمیم‌گیری و تأیید صحیح هویت گوینده نیاز به واحدی است که بتواند به درستی از خروجی‌های الگو، عمل تصمیم‌گیری را انجام دهد. خروجی‌های الگو در واحدهایی با عنوان بردار- x از شبکهٔ عصبی تأخیر زمانی استخراج می‌شوند که توسط تحلیل تفکیک خطی احتمالاتی (فرمول ۶) مشابهت‌سنجی می‌شوند و سرانجام امتیازدهی صورت می‌گیرد. براساس امتیاز به‌دست‌آمده، هویت گوینده تأیید یا رد می‌شود. تحلیل تفکیک خطی از جمله روش‌های آماری است که در یادگیری ماشین و بازشناخت الگو برای پیدا کردن ترکیب خطی خصوصیاتی که به بهترین صورت دو یا چند کلاس از اشیا را از هم جدا می‌کند، استفاده می‌شود [۲۱].

$$f_{plda}\left(W_{target}, W_{test}\right) = \log p\left(W_{target}, W_{test} \mid H_1\right) - \log\left[p\left(W_{target} \mid H_0\right) p\left(W_{test} \mid H_0\right)\right] \quad (۶)$$

در رابطه (۶) H_1، فرض می‌کند که دو گوینده یکسان است و H_0 فرض می‌کند که دو گوینده متفاوت است [۲۱].

به‌منظور ارزیابی روش پیشنهادی برای تأیید هویت گوینده، علاوه بر شبه‌سنج نرخ خطای یکسان (ایی‌ایی‌آر)[1] که در بخش بعدی به‌صورت کامل توضیح داده خواهد شد، معیارهای دیگری نظیر تابع هزینه تصمیم، نسبت نرخ مثبت صحیح (تی‌پی‌آر)[2]، نرخ مثبت غلط (اِف‌پی‌آر)[3]، نرخ پذیرش غلط (اِف‌اِی‌آر)[4] و نرخ رد غلط (اِف‌آرآر)[5] بررسی شده است. هم‌چنین نمودارهای مشخصه عملکرد (آراُسی)[6]، پراکندگی امتیازات و توزیع فراوانی احتمال (پی‌دی‌اِف)[7] ترسیم شده است که هر نمودار به‌طور کامل در بخش سوم شرح داده خواهد شد.

۳. نتایج

۳-۱. نمودارهای مشخصه عملکرد، پراکندگی امتیازات و توزیع فراوانی احتمال

یکی از روش‌های بررسی و ارزیابی عملکرد دسته‌بندی دودویی، نمودار مشخصه عملکرد است. کارایی الگوریتم‌های دسته‌بندهای دودویی[8] معمولاً توسط شاخص‌هایی به نام حساسیت[9] یا بازیابی[10] سنجیده می‌شود. اما در نمودار مشخصه عملکرد هر دو این شاخص‌ها ترکیب شده و به‌صورت یک منحنی نمایش داده می‌شوند. منحنی نمودار مشخصه عملکرد توسط ترسیم

[1] EER; Equal Error Rate
[2] TPR; True Positive Rate
[3] FPR; False Positive Rate
[4] FAR; False Acceptance Rate
[5] FRR; False Rejection Rate
[6] ROC; Receiver Operating Characteristic
[7] PDF; Probability Frequency Distribution
[8] Binary classifier
[9] Sensitivity
[10] Recall

نسبت نرخ مثبت صحیح بر حسب نرخ مثبت غلط (فرمول ۷ و ۸) [۲۱] ایجاد می‌شود. البته باید توجه داشت که آستانه برای این مقادیر، متغیر است. به همین دلیل، یک نمودار پیوسته ایجاد خواهد شد.

$$TPR = TP / P = TP / (TP + FN) \quad (۷)$$

$$FPR = FP / N = FP / (FP + TN) \quad (۸)$$

برای محاسبهٔ نرخ مثبت صحیح و نرخ مثبت غلط ماتریس درهم‌ریختگی[1] طبق جدول ۴ نشان داده شده است.

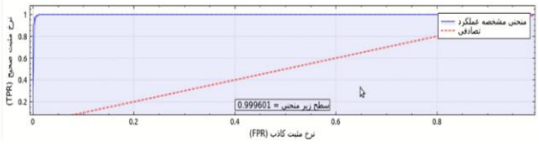
جدول ۴ ماتریس درهم‌ریختگی به منظور محاسبهٔ نرخ مثبت صحیح و نرخ مثبت غلط برای ترسیم نمودار مشخصه عملکرد.

واقعیت / پیش‌بینی	مثبت (پی)	منفی (اِن)
صحیح (تی)	تی‌پی	تی‌اِن
غلط (اِف)	اِف‌پی	اِف‌اِن

مثبت صحیح (تی‌پی)[2] بیانگر ضربه[3] یا شناسایی صحیح است.
منفی صحیح (تی‌اِن)[4] به معنی رد صحیح[5] است. یعنی تشخیص صحیح یک نوفه که ممکن بود به اشتباه علامت تلقی شود.
مثبت غلط (اِف‌پی)[6] همان اخطار غلط[7] است. در این حالت یک نوفه به عنوان علامت شناخته شده است.
منفی غلط (اِف‌اِن)[8] معادل علامت گمشده[9] در آزمون فرض آماری است. به این ترتیب دستگاه دریافت‌کننده، نتوانسته علامت موجود را شناسایی کند.

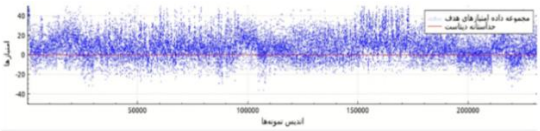
در شکل ۴ نمودار مشخصه عملکرد نشان داده شده است که هر چه نمودار آبی از نمودار قرمز بالاتر باشد، نشان‌دهندهٔ درستی سامانه و هر چه خط آبی پایین‌تر از خط قرمز باشد، خطای سامانه از درستی آن بیش‌تر است. ماتریس درهم‌ریختگی نتایج حاصل از طبقه‌بندی را براساس اطلاعات واقعی موجود، نمایش می‌دهد. معمولاً چنین نمایشی برای الگوریتم‌های یادگیری با ناظر استفاده می‌شود، اگرچه در یادگیری بدون ناظر نیز کاربرد دارد. معمولاً به کاربرد این ماتریس در الگوریتم‌های بدون ناظر، ماتریس تطابق می‌گویند.

[1] Confusion matrix
[2] TP; True Positive
[3] Hit
[4] TN; True Negative
[5] Correct reject
[6] FP; False Positive
[7] False alarm
[8] FN; False Negative
[9] Miss

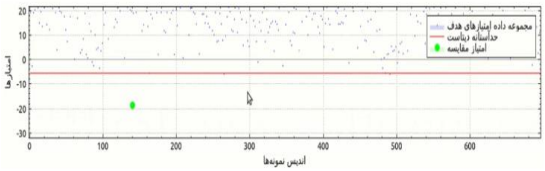


شکل ۴ منحنی نمودار مشخصه عملکرد به‌دست‌آمده از روش پیشنهادی.

هر ستون از ماتریس، نمونه‌ای از مقدار پیش‌بینی‌شده را نشان می‌دهد. در صورتی‌که هر سطر نمونه‌ای واقعی را در بردارد. پراکندگی امتیازات به‌دست‌آمده از مقایسهٔ هر فایل گفتاری در یک آزمایش[10] در شکل ۵ نشان داده شده است. بعد از انتخاب دو فایل صوتی از گوینده(گان)، نقطه سبز در نمودار مجموعه داده، ایجاد می‌شود (شکل ۶) که بیانگر امتیاز مشابهت است. در صورتی‌که نقطه سبز رنگ پایین خط آستانه باشد، دو گوینده متفاوت هستند و در صورتی‌که نقطه مذکور بالای حد آستانه باشد، دو گوینده یکسان هستند.



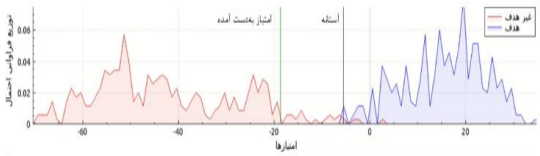
شکل ۵ مشاهدهٔ پراکندگی امتیازات به‌دست‌آمده از کل دادگان معرفی‌شده برای روش پیشنهادی.



شکل ۶ نقطهٔ سبز رنگ، امتیاز به‌دست‌آمده بعد از مقایسهٔ دو فایل صوتی در روش پیشنهادی را نشان می‌دهد. با توجه به این‌که نقطهٔ حاصل، بسیار پایین‌تر از حد آستانه قرار دارد، با احتمال بالایی دو گوینده متفاوت هستند.

نمودار توزیع فراوانی احتمال در شکل ۷ نشان داده شده است. خط سیاه، نشان‌دهندهٔ مقدار آستانه، خط سبز امتیاز به‌دست‌آمده از مقایسه، نمودار آبی چگالی هدف (یعنی گویندگان یکسان) و نمودار قرمز چگالی غیر هدف (یعنی گویندگان غیریکسان) است.

[10] Trial



شکل ۷ نمایش توزیعی نمرات هدف، غیرهدف و آستانه تصمیم‌گیری روش پیشنهادی.

۳-۲. نرخ خطای یکسان

یکی دیگر از شبه‌سنج‌های ارزیابی تأیید هویت گوینده، نرخ خطای یکسان است که با برابر شدن مقادیر نرخ پذیرش غلط و نرخ رد غلط تعریف می‌شود. یعنی با تغییر آستانه، که نرخ پذیرش غلط و نرخ رد غلط (فرمول ۹ و ۱۰) در نقطه‌ای [۲۲-۲۳] برابر شوند. نرخ خطای برابر یک معیار بسیار محبوب برای سامانه‌های تأیید گوینده است. فقط نمرات نرم سامانه خودکار برای محاسبه نرخ خطای برابر لازم هستند [۲۲-۲۳].

$$FAR = \text{تعداد خطاهای اِفاِی}^{2} / \text{تعداد سعی تقلبی}^{1} \quad (۹)$$

$$FRR = \text{تعداد خطاهای اِفآر}^{3} / \text{تعداد سعی قانونی}^{3} \quad (۱۰)$$

۳-۳. تابع هزینه تصمیم

تابع هزینه تصمیم یکی دیگر از معیارهای ارزیابی عملکرد سامانه‌های تأیید هویت گوینده است. تابع هزینه تصمیم، احتمال‌های پیشین اتفاق افتاده گوینده هدف را در نظر می‌گیرد یعنی مشخص می‌کند که گویندهٔ ثبت‌نام‌شده با گویندهٔ در حال آزمایش یکسان است یا خیر که کمینه مقدار آن‌را با کمینه تابع هزینه تصمیم نشان می‌دهند (فرمول ۱۱) [۲۲-۲۳].

$$DCF(x) = C_{MISS} P(x) P_{Target} + C_{FA} P_{FA}(x)(1 - P_{Target}) \quad (۱۱)$$

در رابطه ۱۱ C_{Miss} هزینهٔ خطای Miss/FR، C_{FA} هزینهٔ خطای FA، P_{Target} احتمال پیشین گویندهٔ هدف، $P_{Miss}(\tau)$ احتمال (Miss| هدف، آستانه) و $P_{FA}(\tau)$ احتمال (FA| غیرهدف، آستانه) می‌باشد.

همان‌طورکه در جدول ۵ نشان داده شده است، نرخ خطای برابر در حالتی‌که نرخ نمونه‌برداری ۱۶کیلوهرتز است و از هر دو دادگان وُکس‌سِلِب۱، ۲ تواماً (به‌صورت تمیز) برای آموزش استفاده شده است، بهترین نتیجه معادل با ۳٫۰۹ به‌دست آمده است. زمانی که دادگان وُکس‌سِلِب۱ به تلفنی تبدیل و مجددا آزمایش تکرار شد، نرخ خطای برابر افزایش پیدا کرده و به مقدار ۱۴٫۳۴رسیده است. زمانی که دادگان تلفنی شدهٔ وُکس‌سِلِب۱ به همراه دادگان اصلی مذکور برای آموزش استفاده شده، مقدار خطا از ۱۴٫۳۴ به ۵٫۳۹ کاهش یافته است. هم‌چنین در حالتی‌که برای آموزش از ترکیب دادگان وُکس‌سِلِب۱، ۲ برای آموزش و برای تطبیق از دادگان احساسی ش‌امو[4] و دادگان تلفنی کال‌هوم استفاده شده است (با نرخ نمونه‌برداری هشت کیلوهرتز)، مقدار نرخ خطای برابر نسبت به آموزش با الگوی اصلی و تلفنی وُکس‌سِلِب۱ (با نرخ نمونه‌برداری ۱۶ کیلوهرتز) از ۵٫۳۹ به ۴٫۹۵ کاهش یافته و بهتر شده است.

۴. بحث

در جدول ۶ نتایج حاصل از مقدار نرخ خطای برابر در شرایط نسبتاً تمیز، در حالتی‌که دادگان آموزش فقط وُکس‌سِلِب۱ یا وُکس‌سِلِب۲ و دادگان آزمون وُکس‌سِلِب۱ است، نشان داده شده‌اند. همان‌طورکه مشاهده می‌کنید، در بهترین حالت، نرخ خطای برابر معادل با ۳٫۲۳ و در بدترین حالت ۱۰٫۲ است. آموزش با ترکیب دادگان وُکس‌سِلِب۱، ۲ با دادگان آزمون وُکس‌سِلِب۱ با استفاده از روش پیشنهادی برابر با ۳٫۰۹ شده است که نشان‌دهندهٔ کاهش مقدار نرخ خطای برابر نسبت به الگوهای پیشین است.

۵. نتیجه‌گیری

از جمله روش‌های به روز در تأیید هویت گوینده، استفاده از روش‌های مبتنی بر شبکهٔ عصبی عمیق است که امروزه در مطالعات فراوانی استفاده شده است. اغلب مطالعات صورت‌گرفته در محیط تمیز و فاقد نوفه می‌باشند در صورتی‌که کاربردهای تأیید هویت گوینده اغلب در محیط‌های تلفنی صورت می‌گیرد. بنابراین در این تحقیق پس از پیش‌پردازش دادگان با استفاده از آشکارساز فعالیت صوتی، ضرایب کپسترال بسامد مل برای استخراج

[1] Impostor attempts
[2] FA; False Alarm
[3] Legitimate attempts
[4] ShEmo

جدول ۵ نتایج روش استفاده‌شده روی دادگان آزمون وُکس‌سِلِب۱.

دادگان آموزش	دادگان تطبیق	نرخ نمونه‌برداری	ایی‌ایی‌آر	کمینه دی‌سی‌اِف (پی-هدف= ۰٫۰۱)[1]	کمینه دی‌سی‌اِف (پی-هدف= ۰٫۰۰۱)
Vox۱+ Vox۲	-	۱۶k	۳٫۰۹	۰٫۳۲۶۹	۰٫۵۲۳۸
Vox۱+ Vox۲	-	۸k	۴٫۷۲	۰٫۴۸۷۳	۰/۵۵۹۹
Vox۱+ Vox۲	دادگان تلفنی	۸k	۴٫۹۵	۰٫۵۱۶۸	۰٫۷۰۲۷
Vox۱	-	۱۶k	۷٫۳۶	۰٫۵۷۸۵	۰٫۶۹۴۴
Vox۱	-	۸k	۹٫۰۴	۰٫۶۹۵	۰٫۷۸۸۹
[3]Vox۱+ tell۱	-	۱۶k	۱۴٫۳۴	۰٫۹۱۲۴	۰٫۹۷۷۹
Vox۱+ Vox۱- tell۱	-	۱۶k	۵٫۳۹	۰٫۵۱۸	۰٫۶۳۳۹

[1] Min DCF(p-target=۰٫۰۱)
[2] Min DCF(p-target=۰٫۰۰۱)

[3] منظور از Vox۱+ tell۱ این است که دادگان وُکس‌سِلِب۱ به‌صورت تلفنی تبدیل شده است.

جدول ۶ مقایسهٔ روش‌ها و مدل‌های مختلف بر روی دادگان آزمون وُکس‌سِلِب۱.

الگو	دادگان آموزش	الگوی فِرانت-اِند[1]	اتلاف	ابعاد	ایی‌ایی‌آر
اینتراِسپیچ[2] ۱۷ [۱۳]	وُکس‌سِلِب۱	وی‌جی‌جی-اِم[3]	سافت‌مکس	۱۰۲۴	۱۰٫۲
اینتراِسپیچ ۱۷ [۱۳]	وُکس‌سِلِب۱	وی‌جی‌جی-اِم	سافت‌مکس+ کانترَستیو[4]	۱۰۲۴	۷٫۸
اینتراِسپیچ ۱۸[۱۳]	وُکس‌سِلِب۲	رِس‌نِت- ۳۴	سافت‌مکس+ کانترَستیو	۵۱۲	۴٫۸۳
اینتراِسپیچ ۱۹ [۱۳]	وُکس‌سِلِب۲	رِس‌نِت- ۵۰	سافت‌مکس+ کانترَستیو	۵۱۲	۳٫۹۵
آی‌سی‌اِی‌اِس‌اِس‌پی۱۹[5] [۱۳]	وُکس‌سِلِب۲	تین- رِس‌نِت- ۳۴[6]	سافت‌مکس	۵۱۲	۳٫۵۷
آی‌سی‌اِی‌اِس‌اِس‌پی۱۹ [۱۳]	وُکس‌سِلِب۲	تین- رِس‌نِت- ۳۴	اِی‌اِم- سافت‌مکس[7]	۵۱۲	۳٫۳۲
آی‌سی‌اِی‌اِس‌اِس‌پی۱۹ [۱۳]	وُکس‌سِلِب۲	تین- رِس‌نِت- ۳۴	اِی‌اِم- سافت‌مکس	۵۱۲	۳٫۲۳
آی‌سی‌اِی‌اِس‌اِس‌پی۱۹ (کلیند[8]) [۱۳]	وُکس‌سِلِب۲	تین- رِس‌نِت- ۳۴	سافت‌مکس	۵۱۲	۳٫۲۴
روش پیشنهادی	Vox۱+ Vox۲	تی‌دی‌اِن‌اِن	لُگ- سافت‌مکس[9]	۵۱۲	۳٫۰۹

[1] Front-End
[2] Interspeech
[3] VGG-M
[4] Contrastive
[5] ICASSP19
[6] Thin-ResNet-34
[7] AM-Softmax
[8] Cleaned
[9] Log-Softmax

ویژگی‌های صوتی به‌کار گرفته شده‌اند. ویژگی‌های صوتی مستخرج، به عنوان ورودی به شبکهٔ عصبی تأخیر زمانی داده می‌شوند که خروجی آن بردار- x است. بردارهای- x به بخش تصمیم‌گیری داده می‌شوند تا امتیازی به آن داده شود. اگر امتیاز داده‌شده، بیش‌تر از آستانه باشد، گوینده موردنظر با گویندهٔ از قبل ثبت‌نام‌شده، یکسان است؛ در غیر این صورت، گویندگان متفاوت هستند. سامانه پیشنهادی را ابتدا روی دادگان تمیز آموزش داده که نرخ خطای برابر معادل با ۳٫۰۹ به‌دست آمد. سپس در حالتی‌که از دادگان تلفنی کال‌هوم به عنوان داده تطبیق استفاده شد، ارزیابی صورت گرفت که مقدار نرخ خطای برابر معادل با ۴٫۹۵ به‌دست آمده است. به‌منظور داده‌افزایی، دادگان وُکس‌سِلِب۱ به تلفنی تبدیل شدند و به عنوان دادهٔ آموزشی به سامانه پیشنهادی اعمال گردیدند که نرخ خطای برابر معادل با ۱۴٫۳۴ شد. بنابراین طبق نتایج به‌دست‌آمده، اضافه کردن دادگان تلفنی در حالت تطبیق باعث نتایج بهتری در تأیید هویت گوینده در شرایط تلفنی شده است.

تقدیر و تشکر

در این تحقیق، از جناب آقای داوود سعادتمند، جناب آقای حامد هادی و هم‌چنین آزمایشگاه مرکز فناوری

فضای مجازی برای در اختیار گذاشتن امکانـات پردازشـی، کمال تشکر و قدردانی بعمل می‌آید.

تضاد منافع

نویسندگان این مقاله اعلام می‌کنند که هیچ تضاد منـافعی در مورد محتویات این مقاله ندارند.

فهرست منابع

[1] S. Shahmoradi, S. Bagheri Shouraki, “Evaluation of a novel fuzzy sequential pattern recognition tool (fuzzy elastic matching machine) and its applications in speech and handwriting recognition,” Applied Soft Computing, vol. 62, pp. 315–327, 2018.

[2] Y.C.K., M. Hariharan, R. Ngadiran, A.H. Adom, S. Yaacob, K. Polat, “Hybrid BBO_PSO and higher order spectral features for emotion and stress recognition from natural speech,” Applied Soft Computing, vol. 56, pp. 217–232, 2017.

[3] A.B. Nassif, I. Shahin, I. Attili, M. Azzeh, K. Shaalan, “Speech recognition using deep neural networks: A systematic review,” IEEE Access, vol. 7, pp. 19143–19165, 2019.

[4] A.H. Meftah, H. Mathkour, S. Kerrache, Y.A. Alotaibi, “Speaker identification in different emotional states in Arabic and English,” IEEE Access, vol. 8, pp. 60070–60083, 2020.

[5] J. Neyman, E. Pearson, “On the problem of the most efficient tests of statistical hypotheses,” Philosophical Transacyions of the Royal Society, vol. 231, pp. 289–337, 1933.

[6] J.P. Campbell, W. Shen, W.M. Campbell, R. Schwartz, J. Bonastre, D. Matrouf, “Forensic speaker recognition,” IEEE Signal Processing Magazing, vol. 26, no. 2, pp. 95-103, 2009.

[7] G.S. Morrison, “Distinguishing between science and pseudoscience in forensic acoustics,” in Processing Meetings on Acoustics, pp. 060001, 2013.

[8] P. Rose, “Forensic Speaker Identification,” Boca Raton, FL: CRC Press, 2004.

[9] X. Liu, M. Sahidullah, T. Kinnunen, “A comparative re-assessment of feature extractors for deep speaker embeddings,” in: Processing Interspeech, pp. 3221–3225, 2020.

[10] M. McLaren, L. Ferrer, D. Castan, A. Lawson, “The speakers in the wild (SITW) speaker recognition database,” in: Interspeech, pp. 818-822, 2016.

[11] J. Villalba, N. Chen, D. Snyder, D. Garcia-Romero, A. McCree, G. Sell, J. Borgstrom, L. P. García-Perera, F. Richardson, R. Dehak, P.A. Torres-Carrasquillo, N. Dehak, “State of the art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations,” Computer Speech & Language, vol. 60, 2020.

[12] Z. Bai, X-L. Zhang, “Speaker Recognition Based on Deep Learning: An Overview,” Elsevier, Neural Networks, vol. 140, pp. 65-99, 2021.

[13] A. Nagrani, J.S. Chung, W. Xie, A. Zisserman, “Voxceleb: Large-scale speaker verification in the wild,” Computer Speech & Language, vol. 60, 2020.

[14] A. Canavan, D. Graff, G. Zipperlen, “Callhome American English Speech,” Philadelphia: Linguistic Data Consortium, 1997.

[15] A.F. Martin, C.S. Greenberg, “The NIST 2010 speaker recognition evaluation,” in Processing Interspeech, pp. 2726–2729, 2010.

[16] G.R. Doddington, M.A. Przybocki, A.F. Martin, D.A. Reynolds, “The NIST speaker recognition evaluation–overview, methodology, systems, results, per- spective,” Speech Communication, vol. 31, no. 2–3, pp. 225–254, 2000.

[17] J. Markel, B. Oshika, A. Gray, “Long-term feature averaging for speak- er recognition,” IEEE Transaction of Acoustical Speech Signal Processing, vol. 25, no. 4, pp. 330–337, 1977.

[18] K. Li, E. Wrench, “An approach to text-independent speaker recognition with short utterances,” in Processing IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 8, pp. 555–558, 1983.

[19] H.L. John, H. Taufiq, “Speaker recognitions by humans and machines”, IEEE Signal Processing Magazine, pp. 74-99, 2015.

[20] Wikipedia Contributors. George Zimmerman Available:http://en.wikipedia.org/wiki/George_Zimmerman, 2014.

[21] F. Nolan, T. Oh, “Identical twins, different voices,” International Journal of Speech, Language and the Law, vol. 3, no. 1, pp. 39–49, 1996.

[22] A.M. Martinez, A.C. Kak, “PCA versus LDA,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 228–233, 2001.

[23] N. Dehak, P. Dumouchel, P. Kenny, “Modeling prosodic features with joint factor analysis for speaker verification,” IEEE Transactions Audio, Speech Language Processing, vol. 15, no. 7, pp. 2095–2103, 2007.