بازشناسی آوای فارسی با استفاده از شاخص‌های صوتی و روش‌های جبران‌سازی تنوعاتِ مبتنی بر شبکه‌های عصبی

شقایق رضا[1]، سیّد علی سیّدصالحی[*2]، سیّده زهره سیّدصالحی[3]

[1 و 2] دانشکدهٔ مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، تهران، ایران.

[3] گروه مهندسی پزشکی، دانشکده بهداشت علوم پزشکی تهران، دانشگاه آزاد اسلامی تهران، ایران.

چکیده

شواهد و آزمایش‌های گفتاری نشان می‌دهد که اطلاعات در سیگنال گفتار به‌صورت غیر یک‌نواخت توزیع شده و انسان با تمرکز به نواحی پُر اطلاعات آن قادر است به‌صورت مقاوم گفتار را بازشناسی کند. در این راستا در این پژوهش، یک سامانهٔ بازشناسی آوای فارسی مبتنی بر تمرکز روی بازشناسی مقاوم نواحی پُراطلاعات و مجزای صوتی ارائه شده است. این نواحی شاخص‌های صوتی نامیده می‌شوند. بدین منظور ابتدا برای سیگنال گفتارِ زبان فارسی یک مجموعه از شاخص‌های مناسب صوتی انتخاب شده و به یک شبکهٔ عصبی عمیق آموزش داده شده‌اند، سپس، به‌منظور حذف تنوعات شاخص‌های صوتی، تغییراتی در ساختار مدل و شیوهٔ آموزش آن در چهار طرح مختلف انجام شده است. در طرح اول، از یک شبکهٔ عصبی جداگانه و در طرح دوم از یک ساختار یادگیری چند تکلیفی برای جبران‌سازی غیرخطی تنوعات شاخص‌های صوتی استفاده شده است. در طرح سوم نیز از یک اتصال بازگشتی در لایه‌ی پنهان شبکه برای بازسازی ورودی و در طرح چهارم از یک ساختار مبتنی بر شبکه‌های جاذب‌دار عمیق برای کاهش تنوعات ناخواسته استفاده شده است. در این مقاله آزمایش‌ها روی مجموعه‌دادگانِ گفتاری فارسی "فارس‌دات" انجام شده است و نتایج بازشناسی به‌صورت خطای بازشناسی آوا گزارش شده است. بهترین مدل آموزش‌یافته، یک شبکهٔ عصبی جلوسو با پنج لایهٔ پنهان است. خطای بازشناسی آوای این ساختار روی دادگان آزمون برابر ۲۱/۷۴ درصد به‌دست آمد. همچنین استفاده از چهارطرحِ پالایش تنوعات به ترتیب خطای بازشناسی آوا را به طور مطلق ۰/۳۹، ۰/۵۸، ۰/۴۳ و ۱/۳ درصد کاهش داده است.

واژگان کلیدی: بازشناسی آوا، شاخص‌های صوتی، یادگیری عمیق، بازشناسی مقاوم، پالایش غیرخطی.

Persian Phone Recognition Using Acoustic Landmarks and Neural Network-based variability compensation methods

Shaghayegh Reza[1], Seyyed Ali Seyyedsalehi[2*], Seyyede Zohre Seyyedsalehi[3]

[1,2] Biomedical Engineering Faculty, Amirkabir University of Technology, Tehran, Iran.
[3] Department of Biomedical Engineering, Faculty of Health, Tehran Medical sciences, Islamic Azad, University, Tehran, Iran

Abstract

Speech recognition is a subfield of artificial intelligence that develops technologies to convert speech utterance into transcription. So far, various methods such as hidden Markov models and artificial neural networks have been used to develop speech recognition systems. In most of these systems, the speech signal frames are processed uniformly, while the information is not evenly distributed in all of them. Auditory experiments have also shown that the human brain pays more attention to information-rich areas. By focusing on these areas instead of uniform processing, the brain can more robustly recognize speech in intrinsic and environmental speech variations such as speaker and noise. In

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۳۹۹/۶/۱۷ • تاریخ پذیرش: ۱۴۰۰/۶/۳ • تاریخ انتشار: ۱۴۰۱/۱۲/۲۹ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.19.4.173]

contrast, the performance of most speech recognition systems degrades dramatically in these conditions. Therefore, to boost speech recognition systems' robustness, some researchers have focused on developing speech recognition systems by modeling these informative parts of the speech signal named landmarks. Similarly, in this article, we implemented a landmark-based system to obtain a robust Persian speech recognition system inspired by human brain perception. We also conducted neural networks-based variation compensation methods to boost its performance.

In this article, acoustic landmarks are classified into two categories of events and states with the following definitions. Events are defined as areas of the speech signal in which the spectral characteristics change drastically while their length does not change a lot. The transition areas between some adjacent pairs of phones (phones' borders) are primarily selected as events. States are also defined as areas of the speech signal that spectral characteristics do not change significantly. Here the nuclei of phones are considered as the states. Previous research, linguistic sources, and implementation results have been used to determine the Persian language's appropriate landmarks. Finally, a set of 313 landmarks was selected and used in our acoustic landmarks-based phone recognition system .

The neural network structure used to recognize acoustic landmarks is a feed-forward fully connected structure with ReLU function in its hidden layers and a linear function in its final layer. The number of layers and neurons of this structure has been determined experimentally. The best structure is composed of 5 fully connected layers with 1000 neurons per layer. In this study, instead of considering 313 neurons to express each of the 313 landmarks, a heuristic labeling method is used to reduce the number of output neurons and utilize the shared information between the landmarks. The landmark recognition model slides on the speech feature sequence in the test phase to produce the output landmark sequence. Finally, to convert the obtained landmark sequence to a phone sequence, three rule-based post-processing steps are performed .

Variabilities are among the essential quality degradation sources in speech recognition; therefore, we proposed two approaches to reduce them and boost phone recognition quality in our landmark-based system. To this aim, we have utilized the nonlinear filtering characteristic of neural networks by implementing four neural network schemes. In scheme 1, a feed-forward neural network is first trained to map training landmarks to their corresponding well-recognized samples. Then this structure can act as a nonlinear filter before the landmark recognition block. In scheme 2, a unified structure is simultaneously trained to learn landmark labels and the filtering part. In both of these schemes, we used a recursive loop to increase the chance of attractor manipulation in the structures. In scheme 3, a recursive loop is added to one hidden layer. This loop acts as an input variability simulator and forces the network to recognize the input data and its variations correctly. Finally, in scheme four, a deep attractor neural network-based structure is proposed to shape the structure’s hidden layer components so that it can compensate for variabilities.

The experiments are implemented on a Persian database named Farsdat, and the results are reported using phone error rate (PER) criteria. From every 25-millisecond speech frame, an acoustic feature called LHCB is extracted and combined with delta and delta-delta features of that frame. Every frame's features are concatenated with fourteen adjacent frames and are finally fed to our neural network-based landmark extraction model. The best-trained model obtained the PER of 21.74% on test data. Using scheme one to four, we achieved an absolute PER decrease by 0.39, 0.58, 0.43 and 1.30 percent, respectively. Comparing our landmark-based system's performance with other Persian phone recognition systems shows that this method could perform efficiently as a Persian phone recognition system .

In our future works, we intend to compare our acoustic-based phone recognition system's performance with conventional methods such as CTC in noisy conditions. Besides, it seems that acoustic landmarks can be used to create an alignment of the input speech sequence and the output transcription. Therefore, we will present a combination of CTC-based methods and acoustic landmarks to utilize acoustic landmarks' complementary information. This information might boost the performance and speed of CTC-based speech recognition methods, particularly in low resource languages.

Keywords: Phone Recognition, Acoustic Landmarks, Deep Learning, Robust Recognition, Nonlinear Filtering.

۱- مقدمه

هدف از بازشناسی خودکار گفتار، تبدیل گفتار به نمادهای گسسته‌ای از آواها یا کلمات است. با وجود پیشرفت‌های بسیار در این زمینه، کارایی سامانه‌های بازشناسی خودکار گفتار فعلی در برخورد با تنوعاتی چون تغییر گوینده، نوفه محیط و اعوجاج کانال کاهش می‌یابد [۳۳ و ۵۱]، درحالی‌که مغز انسان اغلب در این شرایط نیز به‌خوبی عمل می‌کند. ازاین‌رو تلاش‌ها برای بهبود مدل‌ها و اصلاح روش‌های آموزش مدل‌های بازشناسی ادامه دارد.

در اغلب سامانه‌های بازشناسی گفتار جدید از مدل مخفی مارکوف[1] [۲۰، ۳۷ و ۴۱]، شبکه‌های عصبی عمیق[2] [۹، ۱۴، ۲۵، ۲۷ و ۵۲] و یا تلفیقی از این دو استفاده می‌شود. به‌عنوان نمونه در پژوهش‌های اخیر در حوزهٔ

[1] Hidden Markov Models (HMM)
[2] Deep Neural Networks (DNN)

[DOI: 10.61186/jsdp.19.4.173]

بازشناسی آوا از شبکه‌های باور عمیق[1] [۱۵]، شبکه‌های کانولوشنی[2] [۵۳] و شبکه‌های حافظه بلندِ کوتاه‌مدت[3] [۴۷] استفاده شده است.

رویکرد غالب در بیشتر سامانه‌های بازشناسی گفتار یادشده، پردازش یک‌نواخت قاب‌های[4] گفتاری است درحالی‌که اطلاعات روی سیگنال گفتار به‌صورت یکنواخت توزیع نشده است [۲۴]. در این سامانه‌ها نواحی پُراطلاعات گفتاری (چون گذر بین آواها) مانند سایر نواحی کم‌اطلاعات پردازش می‌شود. در این صورت با وجود اطلاعات بیشتر، در آموزش به اندازهٔ نواحی کم‌اطلاعات دیده شده و در بازشناسی تمرکز ویژه‌ای روی آن‌ها نمی‌شود.

در مقابل این رویکردِ غالب، پژوهش‌ها نشان داده که مغز انسان روی آن نواحی از سیگنال گفتار که حاوی اطلاعات بیشتری است (شاخص‌های صوتی[5])، تمرکز بیشتری دارد [۱۸ و ۴۰]. شبکه‌های عصبی مغز انسان سامانه‌هایی غیرخطی و دوسویه هستند. پردازش اطلاعات در این نوع شبکه‌ها، بر پایه مجموعه‌ای از جاذب‌های نقطه‌ای مجزا و پویا که یادگیری می‌شوند، انجام می‌پذیرد. از ویژگی‌های مفید شکل‌گیری جاذب‌ها در سامانه‌های غیرخطی ایجاد قدرت بازشناسی مقاوم به تنوعات و نوفه است [۴، ۱۷ و ۱۸]. آزمایش‌های عصب‌شناختی برای درک ارتباط بین حس‌گرها و مغز بیانگر آن است که تغییرات سریع در ورودی حس‌گرها، یک عامل مهم در انتقال اطلاعات اطلاعات حسی به‌صورت وقایعی گسسته به مغز است [۱۸] و [۳۱]. بر طبق این شواهد در [۴۵] ایدهٔ پردازش غیریکنواخت گفتار تحت عنوان تئوری شاخص‌های صوتی مطرح شد. بر اساس این ایده تاکنون سامانه‌های بازشناسی گفتار مختلفی مبتنی بر شاخص‌های صوتی ارائه شده است (مرور شده در بخش دوم) [۶، ۲۳، ۳۰ و ۴۶]. مزایای این روش عبارتند از:

- این رویکرد انطباق بیشتری با نحوهٔ ادراکِ مبتنی بر تمرکزِ انسان بر نواحی دربردارندهٔ شاخص‌های صوتی دارد [۶].
- نواحی حاوی شاخص‌های صوتی از گفتار به دلیل اینکه حاوی اطلاعات بیشتری است، کمتر تحت تأثیر تنوعات قرار گرفته و درنتیجه بازشناسی مبتنی بر این نقاط، بازشناسی مقاوم‌تری خواهد بود [۲۲].
- ظرفیت مدل برای مدل‌سازی قاب‌های کم‌اطلاعات هدر نخواهد رفت و درنتیجه مدل به نحو بهتری روی نواحی مهم آموزش خواهد یافت.

در ادامهٔ کارهای قبلی [۳۸ و ۶]، در راستای ارائهٔ یک سامانهٔ بازشناسی گفتار جامع فارسی مبتنی بر شاخص‌های صوتی، اهداف ما در این پژوهش عبارت است از:

- انتخاب شاخص‌های صوتی مناسب برای سیگنال گفتارِ زبان فارسی و ارائهٔ یک سامانهٔ بازشناسی آوای مبتنی بر تئوری شاخص‌های صوتی.
- بررسی قابلیت شبکه‌های عصبی عمیق برای استخراج شاخص‌های صوتی (تاکنون تنها از HMM [۲۶] و یا شبکه‌های کم‌عمق و دادگان محدود [۵ و ۶] استفاده شده است).
- استفاده از تئوری پالایش غیرخطی[6] در شبکه‌های عصبی برای بازشناسی مقاوم به تنوعات شاخص‌های صوتی و بهبود کیفیت بازشناسی آواها.

ساختار مقاله در ادامه به این شرح است: در بخش دوم مروری بر سوابق استفاده از شاخص‌های صوتی در بازشناسی گفتار انجام، در بخش سوم نحوهٔ انتخاب شاخص‌های صوتی برای سیگنال گفتارِ زبان فارسی بیان و در بخش چهارم نحوهٔ استخراج و استفاده از این شاخص‌ها برای بازشناسی آوا شرح داده شده است؛ سپس در بخش پنجم تغییرات انجام‌شده در ساختار مدل بازشناس برای افزایش مقاومت مدل به تنوعاتِ شاخص‌های صوتی توضیح و در بخش ششم و هفتم نیز نتایج آزمایش‌ها، جمع‌بندی نتایج شرح داده شده است.

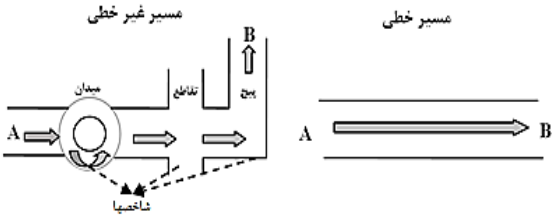
۲- مروری بر سوابق استفاده از شاخص‌های صوتی در بازشناسی گفتار

پدیده‌های غیرخطی را می‌توان با استفاده از شاخص‌های گسستهٔ آنها بیان کرد. مقصود از شاخص‌ها نواحی از آن پدیده است که حاوی اطلاعاتی بیشتر و مفیدتر است. شکل ۱ مثالی ساده را برای مشخص‌شدن منظور از شاخص نشان می‌دهد. در سمت راست این شکل یک مسیر خطی و در سمت چپ یک مسیرغیرخطی نمایش داده شده است. برای بیان مسیرِ غیرخطی بین دو نقطه از واژگانی چون میدان، تقاطع و پیچ استفاده می‌شود. این

[1] Deep Believe Neural Networks (DBNN)
[2] Convolutional Neural Networks (CNN)
[3] Long Short Term Memory Neural Networks (LSTM)
[4] Frame
[5] Acoustic Landmarks
[6] Nonlinear Filtering

[DOI: 10.61186/jsdp.19.4.173]

عناوین شاخص‌های موجود در این مسیر غیرخطی هستند که اطلاعاتی را در زمینه مسیر دربردارند. درصورتی‌که مسیر صاف و خطی باشد، به‌کاربردن این عناوین بی‌معنا است؛ چون شاخصی در مسیر خطی و یکنواخت وجود ندارد و اطلاعات در کل مسیر یکنواخت توزیع شده است.



(شکل-۱): اطلاعات مهم در زمینهٔ یک مسیر غیرخطی توسط شاخص‌های آن بیان می‌شود.

(Figure-1): Important information of a nonlinear path is expressed by its landmarks.

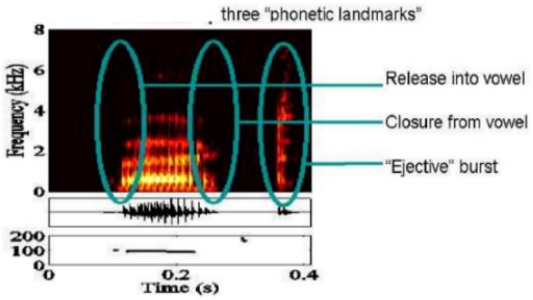
در سیگنال گفتار نیز اطلاعات به‌طور یکنواخت توزیع نشده و در مناطقی از سیگنال گفتار اطلاعات بیشتری وجود دارد. این نواحی پُراطلاعات را شاخص‌های صوتی می‌نامند [۱۲]. شاخص‌های صوتی در پژوهش‌های مختلف متفاوت تعریف شده‌اند. به‌عنوان نمونه:

- در [۴۶]، شاخص‌ها در سه ناحیه گفتاری تعریف شده‌اند: ۱- مناطقی که طیف به‌طورتقریبی ایستاست مثل هستهٔ واکه‌ها[1] و سایشی‌ها[2]، ۲- نواحی گذرای طیف با شیب تند مثل رهش[3] انفجاری‌ها[4] و ۳- نواحی تغییرات تدریجی طیف مثل شبه واکه‌ها[5].
- در [۲۲] ابتدا و انتهای هم‌خوان‌ها[6] و هستهٔ واکه‌ها به‌عنوان شاخص‌های صوتی در نظر گرفته شده‌اند.
- در [۳۰] محل اکسترمم‌ها[7]ی انرژی به‌عنوان شاخص‌های صوتی تعریف شده‌اند.
- برای سیگنال گفتار زبان فارسی در [۵ و ۴۳] از اطلاعات نوع و محل گذر و در [۶] از گذر برخی آواها و ترکیب بعضی سه آواها[8] به‌عنوان شاخص‌های صوتی استفاده شده است.
- در [۳۸]، یک سامانهٔ بازشناسی آوای فارسی با تعریف ۳۱۳ شاخص صوتی در محل هستهٔ آواها و برخی مرزهای آوایی ارائه شد. این پژوهش علاوه‌بر تکمیل و اصلاح آن با ارائهٔ دو روش کاهش تنوعات شاخص‌های صوتی، موجب بهبود نتایج بازشناسی آوا نسبت به نتایج ارائه شده در [۳۸] شده است.

به‌عنوان نمونه در (شکل-۲)، ابتدا و انتهای یک واکه و رهش یک همخوانِ انفجاری به‌عنوان شاخص‌های صوتی در نظر گرفته شده است [۲۲].

تاکنون از شاخص‌های صوتی در سامانه‌های بازشناسی گفتار به روش‌های مختلفی استفاده شده است. به‌عنوان نمونه:

- استفاده از چندین ماشین بردار مرزی (SVM)[9] برای تشخیص انواع شاخص‌های صوتی با استفاده از ویژگی‌های متمایزگر[10] آنها و سپس تلفیق سلسله‌مراتبی اطلاعاتشان برای بازشناسی آواها [۲۶].
- تشخیص شاخص‌های صوتی به‌وسیلهٔ SVMها و استفاده از خروجی آنها برای تعلیم مدل HMM بازشناس گفتار [۱۲].
- استفاده از شبکه‌های عصبی برای بازشناسی آوای مبتنی بر شاخص‌های صوتی [۵ و ۶].
- بررسی ارتباط مناطق حاوی شاخص‌های صوتی با اسپایک‌های سامانهٔ بازشناس گفتار مبتنی بر CTC [۳۵][11].
- استفاده از اطلاعات تکمیلی شاخص‌های صوتی با وزن‌دهی بیشتر قاب‌های حاوی این اطلاعات [۲۳].
- آموزش همزمان شاخص‌های صوتی و برچسب بازشناسی آوا در روش مبتنی بر CTC [۲۴].



(شکل-۲): نمونه‌ای از شاخص‌های صوتی تعریف شده در بخشی از یک دادهٔ گفتاری [۲۲].

(Figure-2): An example of acoustic landmarks in a speech signal segment [22].

بیان این نکته ضروری است که در اغلب مراجع مربوط به سیگنالِ گفتار زبان انگلیسی، هر کدام از ویژگی‌های متمایزگر به‌وسیلهٔ مدل‌های جداگانه‌ای استخراج، سپس اطلاعات این ویژگی‌ها به‌صورت سلسله‌مراتبی برای بازشناسی آواها تلفیق شده‌اند؛ ولی در

[1] Vowels
[2] Fricatives
[3] Release
[4] Explosives
[5] Semi vowels
[6] Consonents
[7] Extremums
[8] Triphones

[9] Support Vector Machines
[10] Distinctive Features
[11] Connectionist temporal classification (CTC)

[DOI: 10.61186/jsdp.19.4.173]

این مقاله بدون نیاز به تعریف ویژگی‌های متمایزگر و استخراج جداگانهٔ آنها، تمامی شاخص‌ها با یک ساختار شبکهٔ عصبی عمیق یاد گرفته می‌شوند.

۳- شاخص‌های صوتی برای سیگنال گفتار زبان فارسی

در این پژوهش، شاخص‌های صوتی به دو دستهٔ واقعه[1] و حالت[2] با تعاریف زیر دسته‌بندی شده‌اند:

- **واقعه:** نواحی‌ای از سیگنال گفتار است که در آن مشخصات طیفی سیگنال تغییرات شدیدی دارد و با تغییر سرعت گفتار، طول آنها زیاد عوض نمی‌شود. در این مقاله نواحی گذر مابین برخی جفت آواهای مجاور به‌عنوان واقعه انتخاب شده‌اند.
- **حالت:** نواحی‌ای از سیگنال گفتار است که مشخصات طیفی تغییرات شدید ندارد. در اینجا هستهٔ آواها به‌عنوان حالت در نظر گرفته شده است.

(جدول-۱): انواع آواهای زبان فارسی.
(Tabel-1): Persian phone types.

نوع برچسب آوایی	تعداد	نماد
واکه‌ها (V)	6	V={â , a, e, o, u, i}
همخوان‌ها (C)	23	C={E ,D}
رهش‌ها (E)	9	E={d, t, b, p, g, q, ?, ĵ, č}
بست‌های تلفیق شده (B)	6	B={1, 2, 3, 4, 5, 6}
همخوان‌های غیر انفجاری (D)	14	D={y, l, m ,n, r, ,k ,f, v, s, z, š, ž, h, x}
دسته آوای A	5	A={s, š, t, n, m}
سکوت (S)	1	S={^}

برای تعریف شاخص‌های صوتی سیگنال گفتار زبان فارسی، ابتدا در جدول-۱ انواع برچسب‌های معمول برای آواهای زبان فارسی نمایش داده شده است. نمادهای به‌کاررفته برای آواهای زبان فارسی با استفاده از مرجع [۲] انتخاب شده است. تعداد آواهای زبان فارسی ۲۹ عدد است، ولی برچسب‌های اختصاص‌یافته به دادگان به‌طورمعمول به‌دلیل برچسب‌زنی مواردی جزئی‌تر چون بست[3] در انفجاری‌ها، بیشتر از تعداد آواها است.

برای تعیین شاخص‌های صوتی مناسب برای سیگنال گفتارِ زبان فارسی، از پژوهش‌های پیشین [۵، ۶ و ۴۳]، منابع زبان‌شناسی [۲] و همچنین خروجی‌های مدل با تعداد شاخص‌های صوتی بیشتر استفاده شده است. در نهایت مجموعهٔ ۳۱۳ تایی شاخص‌های صوتی انتخاب شده (جدول۲) با توجه به محدودیت‌های زیر تعیین شده است:

- کلِّ رهش به عنوان یک شاخص صوتی در نظر گرفته شده است.
- در محل گذر بین بست‌ها و رهش متناظرشان، واقعه در نظر گرفته نشده است.
- در گفتار فارسی گذرهای V–V، V–E، S–E و S–V وجود ندارد ("–" نماد گذر است).
- بست‌های مشابه در انفجاری‌ها تلفیق شده است، زیرا بست انفجاری‌ها در برخی جفت آواها بسیار یکسان‌اند. این موارد شامل بست‌های زوج‌های آوایی (b و p)، (d و t) و (ĵ و č) است.
- اغلب گذرهای C–S خوب آموزش نمی‌بیند و به این دلیل برای عدم افزایش تناقض در یادگیری، تعدادی از آنها حذف شداند و تنها گذرهای مجموعه A+S آموزش می‌یابد. این مجموعه به‌طور تجربی به‌دست آمده است.
-

(جدول-۲): شاخص‌های صوتی انتخاب‌شده.
(Tabel-2): Selected acoustic landmarks.

نوع شاخص صوتی	تعداد شاخص
V–{D,B}	6×20
V–S	6×1
C–V	23×6
S–D	1×14
A–S	5×1
هسته‌ی {V,S,C}	30
مجموع شاخص‌های صوتی	313

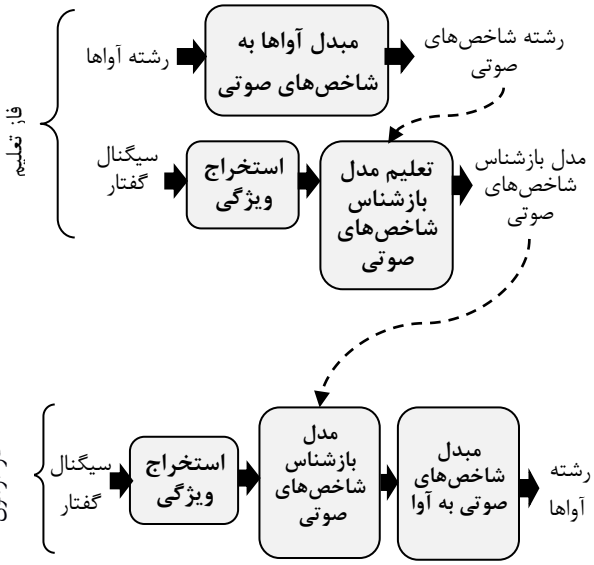
۴- سامانهٔ بازشناسی آوای مبتنی بر شاخص‌های صوتی

مراحل آموزش و آزمون سامانهٔ بازشناسی آوای مبتنی بر شاخص‌های صوتی در (شکل-۳) نمایش داده شده است. ابتدا برچسب‌های آوایی دادگان به برچسب متناسب با شاخص‌های صوتی تبدیل می‌شود (زیربخش ۴-۱). سپس پس از استخراج ویژگی، مدل شبکهٔ عصبی برای نگاشت ورودی به برچسب شاخص‌های صوتی آموزش می‌بیند (زیربخش ۴-۲). در مرحلهٔ آزمون نیز

[1] Event
[2] State
[3] Closure

[DOI: 10.61186/jsdp.19.4.173]

ابتدا بردار ویژگی استخراج شده و با استفاده از مدل آموزش یافته در مرحلهٔ آموزش، رشته شاخص‌های صوتی از دادهٔ آزمون استخراج می‌شود. زنجیرهٔ شاخص‌های صوتی با اعمال پردازش‌هایی که در زیربخش ۴-۳ شرح داده شده است، به زنجیرهٔ آوایی تبدیل می‌شود.



(شکل-۳): سامانه‌ی بازشناسی آوای مبتنی بر شاخص‌های صوتی.

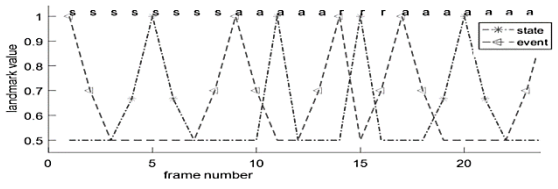
(Figure-3): Block diagram of the landmark based phone recognition system.

۴-۱- برچسب‌زنی شاخص‌های صوتی

برای آموزش مدل بازشناس مبتنی بر شاخص‌های صوتی، کلیّهٔ دادگان موجود باید با برچسب‌هایی که شاخص بودن و نبودن هر قاب گفتاری و نوع شاخص را تعیین می‌کند، برچسب‌زنی شوند. در این مقاله از برچسب‌های نرم استفاده شده است. مقصود ما از برچسب نرم، اختصاص خروجی به‌صورت یک تابع مثلثی در محل شاخص‌ها است به‌طوری‌که در محل شاخص مقدار آن برابر یک و با دورشدن از محل شاخص به‌صورت متقارن در دو طرف، مقدار آن به‌تدریج به صفر برسد. این شیوه برچسب‌زنی دو مزیت دارد:

- در صورت وجود خطای برچسب‌های آوایی در حد یک یا دو قاب، شاخص مورد نظر از آموزش حذف نمی‌شود، بلکه با مقدار خروجی کمتری از یک در آموزش مشارکت می‌کند.
- برای آواهای طولانی مثل واکه‌ها، تعیین محل مشخصی به‌عنوان هستهٔ واج سخت است و بهتر است، چندین قابِ حول مرکز به عنوان هستهٔ آوا در نظر گرفته شود.

در (شکل۴) نحوهٔ اختصاص برچسب نرم برای شاخص‌های صوتی بخشی از یک داده گفتاری (رشته آوای "sara") نمایش داده شده است. منحنی (-*-) مربوط به امتیاز حالت‌ها و منحنی (-◂-) مربوط به امتیاز وقایع است. قاب‌های با امتیاز بالاتر از ۰.۵ به‌عنوان شاخص آموزش داده می‌شوند. در این شیوهٔ برچسب‌زنی برای آواهای با طول کمتر از سه قاب، حالت در نظر گرفته نشده است، زیرا درعمل این آواها در دو گذر قبل و بعدشان به‌صورت واقعه مشارکت دارند.



(شکل-۴): برچسب‌زنی نرمِ شاخص‌های صوتی برای بخشی از یک دادهٔ آموزش.

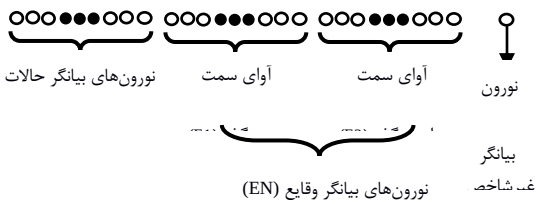
(Figure-4): Soft labeling of acoustic landmarks for a segment of a training data.

در این پژوهش به جای درنظرگرفتن ۳۱۳ نورون برای بیان طبقهٔ هر یک از ۳۱۳ عدد شاخص صوتی در خروجی (بیان‌شده در جدول-۲) ، از شیوهٔ برچسب‌زنی دیگری استفاده شده تا ضمن کاهش تعداد نورون‌های خروجی، از اطلاعات مشترک بین شاخص‌ها نیز استفاده شود. در این راستا، خروجی شبکه ۱۰۳تایی و شامل چهار دسته نورون به تعداد ۳۰، ۳۶، ۳۶ و ۱ است. عدد ۳۰ بیان‌گر مجموعه {C,V,S} برای بیان حالات است و عدد ۳۶ بیان‌گر مجموعه {C,S,V,B} برای بیان آواهای مشارکت‌کننده در وقایع است. دستهٔ نخست این نورون‌ها برای بیان حالت‌ها (SN) و دسته دوم و سوم برای بیان وقایع (EN) مورد استفاده قرار می‌گیرد (شکل-۵). وقتی شاخص صوتی حالت است، خروجی مطلوبِ یکی از ۳۰ نورون اول مقدار دارد و مقدار سایر نورون‌ها صفر است؛ همچنین وقتی شاخص صوتی واقعه است، خروجی مطلوبِ دو نورون از دو دسته نورون ۳۶ تایی مقدار دارند و مقدار سایر نورون‌ها صفر است. این دو نورون مربوط به دو آوای سمت راست و چپ گذر در واقعهٔ مورد نظر است. قاب‌هایی که حالت یا واقعه نیستند، غیر شاخص صوتی (NN) محسوب شده و تنها تک نورون آخر مربوط به غیر شاخص‌بودن در این حالت برابر یک قرار داده می‌شود.

مزیت این شیوهٔ برچسب‌دهی خروجی این است که خروجی مطلوب مشابهی به وقایع مشابه، تخصیص داده می‌شود. برای مثال گذرهای b–â و b–i به دلیل اشتراک آوای 'b'، کد خروجی مشابهی در مقایسه با وقایع

[DOI: 10.61186/jsdp.19.4.173]

بدون اشتراک آوا با این دو گذر دارند.

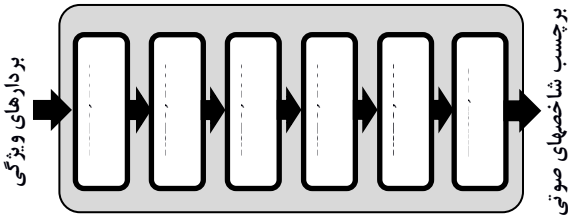


(شکل-۵): نورون‌های خروجی بیان‌گر شاخص‌های صوتی.
(Figure-5): Output neurons representing acoustic landmarks.

۴-۲- مدل بازشناسِ شاخص‌های صوتی

ساختار شبکهٔ عصبی به کار رفته برای بازشناسی شاخص‌های صوتی یک ساختار جلوسو[1] است که شامل تعدادی لایهٔ پنهان تمام متصل (FC)[2] با تابع فعالیت یک‌سوساز خطی[3] (ReLU) در لایه‌های پنهان و خطی در لایهٔ آخر است (شکل-۶). با توجه به توضیحاتی که در مورد نحوه‌ی بیان خروجی در زیر بخش ۴-۱ بیان شد، برای بیان وقایع صوتی دو نورون از نورون‌های خروجی مقدار یک خواهد داشت. از این رو برای لایه‌ی خروجی شبکه، تابع خطی استفاده شده است.

تعداد لایه‌ها و نورون‌های این ساختار به‌صورت تجربی تعیین شده است. بهترین ساختار بدست آمده یک ساختار ۵ لایه پنهان با تعداد نورون‌های هزار در هر لایه است. در شکل (۶) ساختار مدل، تابع غیرخطی و تعداد نورون‌های هر لایه نمایش داده شده است.



(شکل-۶): مدل شبکهٔ عصبی بازشناس شاخص‌های صوتی.
(Figure-6): Neural network model for acoustic landmark recognition.

۴-۳- نحوهٔ تبدیل رشتهٔ شاخص‌های صوتی به رشتهٔ آواها

در مرحلهٔ آزمون، مدل بازشناس آوای مبتنی بر شاخص‌های صوتی روی گفتار می‌لغزد و برای هر قاب، یک بردار خروجی ۱۰۳ بعدی تولید می‌کند. برای تبدیل خروجی مدل به رشتهٔ آواها، سه گام پردازش انجام می‌شود. این سه گام عبارت‌اند از: ۱- تبدیل رشته بردارهای ویژگی قاب‌ها به رشته‌شاخص‌های صوتی ۲- پالایش و تلفیق اطلاعات رشته شاخص‌های صوتی برای تولید رشته آواها ۳- پالایش رشته آواها. در ادامه این سه مرحله شرح داده شده است.

گام نخست: در گام نخست برای پیداکردن محل شاخص‌های صوتی، تصمیم‌گیری اولیه روی خروجی مدل بازشناس شاخص‌های صوتی ({ON={SN,EN,NN) انجام شده و به قاب‌های گفتاری برچسب حالت ('s')، واقعه ('e') و یا غیرشاخص ('n') داده می‌شود؛ بدین منظور محل بیشینه خروجی اگر مربوط به حالت باشد و مقدار آن از سطح آستانهٔ [4] خاصی (thr1) بیشتر باشد، حالت برچسب زده می‌شود. همچنین در صورتی که این مقدار بیشینه، مربوط به واقعه باشد و بیشینه هر دو دسته نورون‌های مربوط به آن (E1 و E2)، از سطح آستانه خاصی (thr2) بیشتر باشد، واقعه در نظر گرفته می‌شود. در غیر این صورت، غیرشاخص بر چسب زده می‌شود. میزان سطوح آستانهٔ thr1 و thr2 در آزمایش‌ها به ترتیب برابر ۰.۲ و ۰.۱ در نظر گرفته شده است. برای تنظیم مقادیر این پارامترها از دادگان توسعه استفاده شده است. با حذف موارد برچسب خورده به‌صورت غیر شاخص ('n')، رشته‌قاب‌های ورودی به رشته‌شاخص‌های صوتی با طولی کمتر از تعداد قاب‌ها تبدیل می‌شود.

الگوریتم-۱: روش تصمیم‌گیری در مورد نوع شاخص صوتی هر فریم.
Algorithm-1: Decision making method of each frame's landmark type.

- if max(ON) = max(SN) & max(SN) > thr1
  Then: 's': the frame is a state
- else if max(ON) = max{EN} & max(E1) & max(E2)>thr2
  Then: 'e': the frame is an event
- else: 'n': the frame is not a landmark

گام دوم: در گام دوم شاخص‌های صوتی نامعتبر پالایش شده و اطلاعات حالت‌ها و وقایع تلفیق می‌شود تا به رشتهٔ آواها تبدیل شود. بدین منظور مراحل زیر بر رشتهٔ شاخص‌های صوتی اعمال می‌شود. در اینجا حالت‌ها با تک نویسه و وقایع با جفت نویسه (برای مثال–â b) نمایش داده شده‌اند.

- تمامی شاخص‌های صوتی تولیدشده که جزء مجموعهٔ شاخص‌های صوتی تعریف‌شده در جدول-۲ نباشند،

[1] Feedforward
[2] Fully Connected
[3] Rectified Linear Unit
[4] Threshold

[DOI: 10.61186/jsdp.19.4.173]

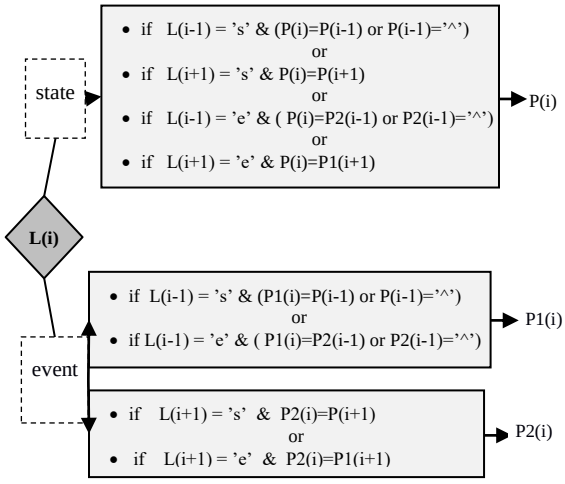
حذف می‌شوند.

- شاخص‌های صوتی تکراری حذف می‌شوند.
  (مثال: ’b b b b–â b–â â â â’→’b b–â â’)
- درصورتی‌که بست شاخص صوتی قبلی و رهش شاخص صوتی بعدی منطبق نباشد، حذف می‌شود. برای مثال حالت ’r‘ به دلیل عدم تطابق بست واقعه ’1^‘ با آن، حذف می‌شود.
  (مثال ’^1 r b–â ’ → ’^1 b–â‘)

برای تلفیق اطلاعات وقایع و حالت‌ها و تبدیل آنها به رشتهٔ آواها، شاخص‌های صوتی متوالی مورد بررسی قرار گرفته و در صورت انطباق یک آوا در دو شاخص متوالی، آن آوا به عنوان آوای درست، در نظر گرفته می‌شود. برای تبیین بیشتر این موضوع، در صورتی که L(i) برچسب شاخص iام (حالت ’s‘ یا واقعه ’e‘) باشد و P(i) آوای مربوط به این شاخص در صورت حالت‌بودن و P1(i) و P2(i) به‌ترتیب آواهای سمت راست و چپ‌گذر در صورت واقعه‌بودن باشند، این قاعده به‌صورت زیر اعمال می‌شود و خروجی آن آوای بازشناسی شده (OutPhone) است. نماد ’^‘ مربوط به سکوت است.

الگوریتم ۲: روش تلفیق اطلاعات شاخص‌های صوتی برای تبدیل رشتهٔ شاخص‌های صوتی به رشته آوا.

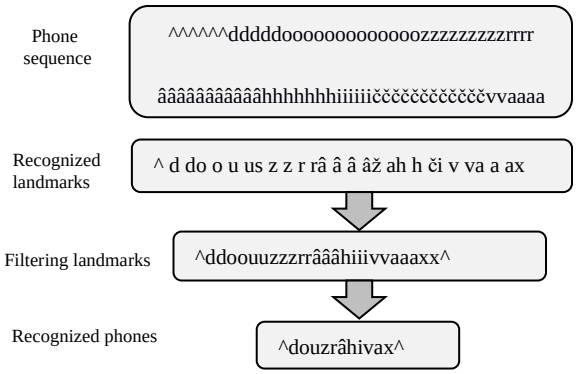
Algorithm-2: Acoustic landmarks' information combination method to transform a landmark sequence to a phone sequence.



گام سوم: در مرحلهٔ سوم با اعمال قواعد زیر، رشته آوای پالایش شدهٔ نهایی تولید می‌شود:

- اگر بین چندین سکوت، بست باشد، سکوت در نظر گرفته می‌شود. (مثال: ’^^^^^‘ → ’^^1^^‘)
- اگر بعد از بست، رهش مربوط به آن نبوده و قبل آن سکوت باشد، سکوت در نظر گرفته می‌شود. به عنوان مثال رهش’1‘ با آوای ’r‘ بعدی انطباق ندارد.
  (مثال: ’^r‘ → ’^1r‘)
- آواهای بین سکوت طولانی حذف می‌شوند. (مثال: ’^^^b^^^‘ →’^^^^^^‘)
- سکوت کوتاهِ بین آواها حذف می‌شود. (مثال: ’bbbb^bbb‘ → ’bbbbbbb‘)
- آواهای تکراری حذف می‌شوند. (مثال: ’bbbâââââ‘ → ’bâ‘)
- درنهایت بست‌ها حذف می‌شوند.
  (مثال: ’^tâ‘ → ’^^^^11ttâââ‘)

در شکل-۷، برای یک جملهٔ نمونه، خروجی هر مرحله نمایش داده شده است.



(شکل-۷): یک مثال از اعمال سه گامِ پردازش برای استخراج رشتهٔ آواهای خروجی از رشته شاخص‌های صوتی.

(Figure-7): An example of the three step process to extract phone sequence from landmark sequence.

۵- پالایش تنوعات شاخص‌های صوتی

در این بخش ابتدا به روش‌های برخورد با تنوعات ناخواسته و منابع این تنوعات در بازشناسی گفتار پرداخته شده، سپس برای کاهش این تنوعات ناخواسته در سامانهٔ بازشناسی گفتار مبتنی بر شاخص‌های صوتی، سه ساختار مبتنی بر شبکه‌های عصبی از پژوهش‌های پیشین معرفی شده و همچنین یک ساختار مبتنی بر شبکه‌های عصبی عمیق جاذب‌دار (DANN)[1] نیز پیشنهاد شده است.

۵-۱- چالش تنوعات ناخواسته در مسائل بازشناسی

به طور کلی مسئله‌ی تنوعات ناخواسته یکی از چالش‌های مهم در مسائل بازشناسی الگو است. منابع این تنوعات دو دسته است: ذاتی و محیطی. مقصود از تنوعات ذاتی، تنوعاتی است که مربوط به پیچیدگی‌های مسئلهٔ بازشناسی است. برای مثال سیگنال گفتار انسان به‌طورذاتی به مشخصات فیزیکی سامانهٔ تولید گفتار

[1] Deep Attractor Neural Networks (DANN)

[DOI: 10.61186/jsdp.19.4.173]

گویندهٔ آن وابسته است. درنتیجه تنوع گوینده یک منبع تنوع ذاتی سیگنال گفتار است. در مقابل مقصود از تنوعات محیطی تنوعاتی است که از ناشی از عوامل بیرونی است. برای مثال نوفه‌های حاصل از کانال تلفن یک تنوع مزاحم محیطی است که بر کیفیت سیگنال گفتار تاثیر می‌گذارد. از آنجا که در این پژوهش هدف بازشناسی گفتار میکروفونی است و دادگان مورد استفاده از میکروفون‌هایی با کیفیت به‌نسبه یکسان و در محیط تاحدودی سکوت ضبط شده است، تمرکز این پژوهش بر کاهش تنوعات ذاتی گفتار است.

در پژوهش‌های مختلف تا کنون رویکردهای مختلفی برای کاهش اثر تنوعات ناخواسته و بهبود مدل‌های بازشناسی به کار رفته است. این رویکردها را می‌توان به‌صورت زیر دسته‌بندی کرد:

- تعلیم دادگان متنوع به مدل بازشناسی: در این رویکرد تا حد ممکن با تأمین داده بیشتر یا ساخت دادگان مصنوعی، سعی می‌شود تا انواع تنوعات ممکن به مدل آموزش داده شود تا مدل بتواند برای کارایی مناسب در دادگان متنوع تقویت شود [۳۴].
- تطبیق مدل بازشناسی به تنوعی خاص: در این رویکرد مدل از پیش تعلیم یافته به تنوع مورد نظر تطبیق می‌یابد. برای این منظور اغلب تمام یا بخشی از پارامترهای مدل برای آن کاربرد خاص اصلاح می‌شود مثل تطبیق مدل بازشناسی گفتار به گوینده خاص [۴۲].
- جبران‌سازی تنوعات پیش از بازشناسی: در این دسته روش‌ها پیش از آموزش مدل بازشناسی، تنوعات ناخواسته تخمین زده شده و حذف شود [۱۶].
- مدل بازشناسی مقاوم به تنوعات: در این رویکرد مدل به نحوی آموزش می‌بیند که در ساختار خود به نحو مناسبی تنوعات را حذف کرده و بازشناسی را همزمان انجام دهد [۱۷].

۲-۵- تنوعات ناخواسته در سامانهٔ مبتنی بر شاخص‌های صوتی

در زیربخش‌های قبل شاخص‌های صوتی به‌عنوان مناطقی با اطلاعات مفید در بازشناسی گفتار معرفی شده و یک سامانهٔ بازشناسی آوای مبتنی بر استخراج شاخص‌های صوتی ارائه شد. به‌دلیل ماهیت گذرا و کوتاه وقایع و انتخاب مناطق ایستان برای حالات، شاخص‌های صوتی تعریف‌شده به تنوعات سرعت بیان به‌نسبه مقاوم‌اند. با این وجود، منابع تنوع دیگری چون گوینده، جنسیت و نوفه محیطی همچنان موجود بوده و می‌توانند دقت بازشناسی شاخص‌های صوتی و درنتیجه دقت بازشناسی گفتار مبتنی بر آن را کاهش دهند.

با توجه به رویکردهای مختلف یادشده برای کاهش تنوعات ناخواسته (زیر بخش ۵-۱)، رویکرد نخست به‌دلیل عدم دسترسی به دادگان متنوع برچسب‌دار فارسی مورد استفاده قرار نگرفته است. همچنین با توجه به اینکه در این پژوهش هدف استفاده از سامانه‌ی بازشناسی گفتار برای گوینده یا جنسیت خاصی نیست و دامنه دادگان موجود نیز بدون تغییر است (میکروفونی و از پایگاه داده یکسانی است)، رویکرد دوم نیز مورد نظر نیست؛ بنابراین در این پژوهش بیشتر بر روش‌های مبتنی بر جبران‌سازی تنوعات پیش از بازشناسی و یا بر مقاوم‌سازی ساختار مدل به تنوعات (دسته رویکرد سوم و چهارم) تمرکز شده است. در زیر بخش بعد روش‌های پیاده‌سازی‌شده در این پژوهش معرفی شده است.

۳-۵- روش‌های جبران‌سازی تنوعات شاخص‌های صوتی با استفاده از شبکه‌های عصبی

در ادامه چهار طرح برای جبران‌سازی تنوعات شاخص‌های صوتی با استفاده از شبکه‌های عصبی معرفی شده است. سه طرح ابتدایی از پژوهش‌های پیشین و طرح چهارم از پیشنهادهای این پژوهش است.

۱-۳-۵- طرح نخست: پالایش غیر خطی تنوعات با استفاده از شبکهٔ عصبی خود کُدکنندهٔ پالایش‌گر نوفه (DAE)[1]

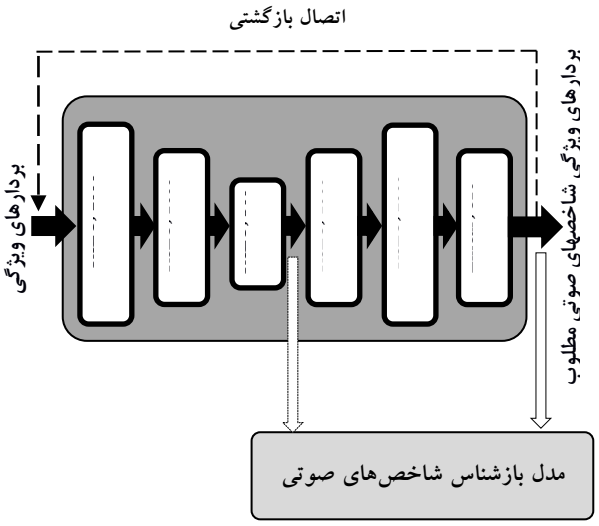
در این روش پیش از بازشناسی، ورودی از یک شبکهٔ حذف‌کنندهٔ تنوعات عبور داده‌شده و سپس خروجی این شبکه به شبکهٔ بازشناسی داده می‌شود (شکل ۸). شبکهٔ عصبی حذف‌کنندهٔ تنوعات، به نحوی تعلیم می‌بینند که نمونه‌های متنوع هر شاخص صوتی را به نمونهٔ بهینهٔ متناظرش نگاشت کند. این ساختار در واقع یک ساختار خودکُدکنندهٔ پالایش‌گر نوفه است.

تاکنون از مفاهیمی از این دست در پژوهش‌های دیگری استفاده شده است. برای مثال در [۴۹ و ۵۰] از نگاشت دادگان نوفه‌شدهٔ ارقام دست‌نوشتار به نمونه عاری از نوفه استفاده شده است. همچنین در [۱۶] به‌طور مشابه از یک شبکهٔ دیگر انجمنی برای تبدیل حروف

[1] Denoising Autoencoder Neural Networks

[DOI: 10.61186/jsdp.19.4.173]

دست‌نوشتار به نمونه‌های دست‌نوشتار مرجع (با دست‌خط خوب) استفاده شده است.



(شکل-۸): طرح ۱- جبران‌سازی تنوعات شاخص‌های صوتی با استفاده از ساختار DAN.

(Figure-8): Scheme 1- Landmark variabilty compensation method using a DAN structure.

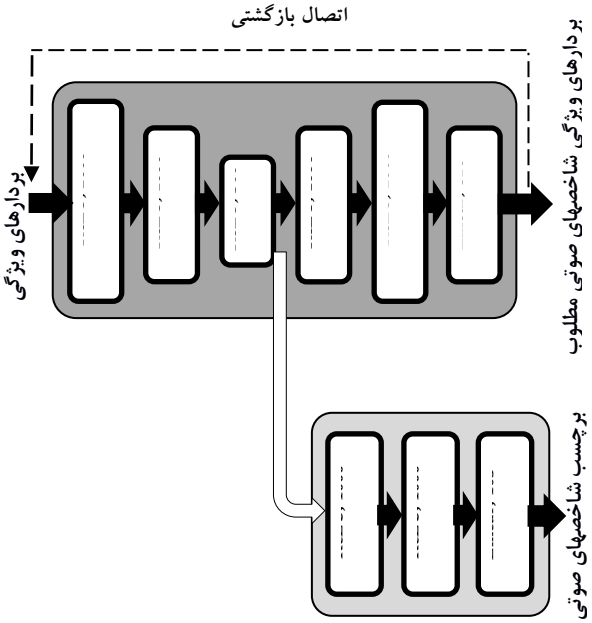
تعیین نمونه‌های بهینه در برخی کاربردها مانند دست نوشتار می‌تواند معنای مشخصی چون دست‌نوشتار خوش‌خط داشته باشد [۱۶]. در بازشناسی گفتار می‌توان گفتار واضح و بدون لهجه را معیاری برای تعیین نماینده قرار داد؛ ولی از آنجا که در دادگان موجود، بیان تمامی شاخص‌های صوتی توسط همه افراد انجام نشده است، امکان تعیین نماینده از این طریق وجود ندارد. از این رو در این پژوهش برای تعیین شاخص‌های صوتی بهینه از مدل بازشناسی شاخص‌های صوتی تعلیم یافته روی دادگان تعلیم (شکل-۶) استفاده شده است. بدین منظور نمونه‌ای از هر شاخص صوتی که به خروجی مطلوب آن شاخص صوتی نزدیک‌تر است، به‌عنوان شاخص‌های صوتی بهینه انتخاب شده است.

۱-۱-۳-۵- اتصال بازگشتی در ساختار DAE

به‌طورکلی ساختارهای DAE برای مثال شکل خاصی از شبکه‌های خودکُدکننده در تئوری قادرند نمونه‌های نوفه‌ای را به نمونه‌های فاقد نوفه تبدیل کرده و در محل نمونه‌های تمیز جاذب در مدل شکل دهند. بنابراین در صورت شکل‌گیری جاذب مناسب می‌توانند مشابه شبکه‌های عصبی خودکُدکننده عمل کرده و با ایجاد اتصال از خروجی به ورودی و اعمال چند بارۀنگاشت، نمونه‌های نوفه‌ای را بیشتر پالایش کنند. با این وجود به دلیل پیچیدگی دادگان، کمبود دادگان، محدودیت ساختار و محدودیت تعلیم مدل، ممکن است امکان ایجاد جاذب مناسب و یا بستر جذب مناسب ممکن نباشد [۴۴]. تا کنون روش‌هایی چرخه‌ای[1] و نمونه‌برداری[2] برای شکل‌گیری مناسب جاذب‌ها در ساختارهای خود کُدکننده و خود کُد کننده‌ی پالایشگر نوفه پیشنهاد شده است [۸، ۲۱ و ۳۹]. در این پژوهشنیز این روش‌ها مورد استفاده قرار گرفته است.

۲-۱-۳-۵- محل خروجی ساختار DAE

هدف از به کار گیری ساختار DAE پالایش تنوعات است و نیازی به بازسازی خروجی وجود ندارد. ز این رو همانطور که در شکل-۸ نمایش داده شده، می‌توان از خروجی لایه‌ی پنهان این ساختار به عنوان ویژگی‌های جدید پالایش شده به جای خروجی آن استفاده کرد. در آزمایش‌های انجام شده در این پژوهشهر دو این خروجی‌ها بررسی شده است.



(شکل-۹): طرح ۲ جبران‌سازی- آموزش توأم برچسب شاخص‌های صوتی و پالایش تنوعات شاخص‌های صوتی.

(Figure-9): Scheme 2- Learning landmark labels and landmark variability compensation in a unified model.

۲-۳-۵- طرح دوم: آموزش توأم برچسب شاخص‌های صوتی و پالایش تنوعات شاخص‌های صوتی توسط مدل مبتنی بر DAN

در طرح نخست دو شبکه به‌صورت مجزا برای حذف تنوعات و بازشناسی مورد آموزش قرار گرفت. این رویکرد مشکل مشخصی دارد: بهینه‌سازی یک ساختار مستقل از دیگری انجام می‌شود و ممکن است، این ساختارهایی که مجزا تعلیم یافته‌اند، برای کاربرد نهایی هم‌سو و بهینه

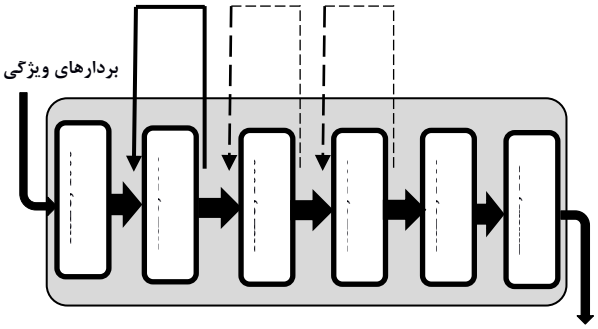
[1] Cyclic
[2] Sampling

[DOI: 10.61186/jsdp.19.4.173]

عمل نکنند. برای رفع این مشکل، در طرح دوم به‌صورت هم‌زمان بازشناسی شاخص‌های صوتی و فیلترکردن نمونه‌های متنوع هر شاخص صوتی انجام و وزن‌های بخش جبران‌سازی و بازشناسی با هم بهینه می‌شود. در شکل ۹ این طرح نمایش داده شده است. به‌طور مشابه در این طرح نیز می‌توان از اتصال بازگشتی برای ایجاد قابلیت جاذب در ساختار مدل استفاده کرد.

۳-۳-۵- طرح سوم: استفاده از اتصال بازگشتی در لایه‌های پنهان برای جبران‌سازی تنوعات ورودی

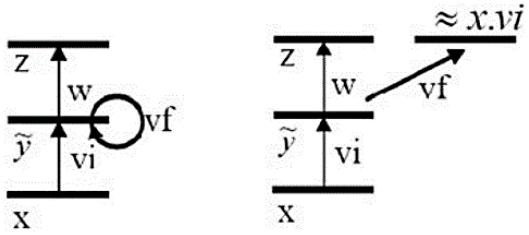
در طرح سوم (شکل-۱۰)، بر خلاف طرح نخست و دوم نیازی به تعیین شاخص‌های صوتی مطلوب نیست. در این طرح اتصالی بازگشتی در یک یا چند لایهٔ پنهان شبکه اعمال می‌شود و عملکرد این اتصال موجب افزایش قدرت مقاومت به تنوعات ورودی می‌شود. استفاده از چنین اتصالاتی ابتدا در [۳۶] برای پیداکردن دادگان مفقود ورودی پیشنهاد شد، سپس این ایده با اصلاح و تغییر شیوهٔ تعلیم در [۱۷] برای حذف نوفه از گفتار بهبود داده شد و در [۱۰] نیز به طور موفق برای تخمین ساختار دوم پروتئین به کار رفت.



(شکل-۱۰): طرح ۳ جبران‌سازی- استفاده از اتصالات بازگشتی در لایه‌های پنهان برای جبران‌سازی تنوعات ورودی.
(Figure-10): Scheme 3- Using recurrent connections in hidden layers to compensate input variabilities.

نحوهٔ عملکرد این اتصال بازگشتی در یک شبکهٔ یک لایهٔ پنهان در شکل-۱۱ با نمایش نقش این اتصال به‌صورت باز شده نمایش داده شده است. این اتصال بازگشتی در واقع مسئله‌ی یادگیری طبقه‌بندی را به یک مسئله‌ی یادگیری توام طبقه‌بندی و بازسازی ورودی تبدیل می‌کند.

جزئیات نحوهٔ تعلیم این اتصال بازگشتی در [۱۷] شرح داده شده است. با توجه به شیوه‌ی تعلیم این اتصالات بازگشتی، طرح ارائه شده در شکل (۱۰) موجب می‌شود که مؤلفه‌های لایه‌های پنهان شبکه در راستای بازسازی شاخص‌های صوتی ورودی شکل گیرد و هم‌زمان شاخص‌های صوتی هم‌طبقه به‌درستی دسته‌بندی شوند؛ درنتیجه در صورت تعلیم مناسب این ساختار، مؤلفه‌های لایه‌های پنهان به‌نحوی شکل می‌گیرد که تنوعات زائد را در راستای دسته‌بندی بهتر شاخص‌های صوتی حذف کنند.



(شکل-۱۱): ساختار یک شبکهٔ عصبی بازگشتی با اتصال بازگشتی در لایهٔ پنهان [۱۷].
(Figure-9): Structure of the recurrent neural network with the recurrent connection to the hidden layer [17].

۴-۳-۵- طرح چهارم: استفاده از ساختار DANN برای شکل‌دهی مؤلفه‌های لایه‌های پنهان برای جبران‌سازی تنوعات

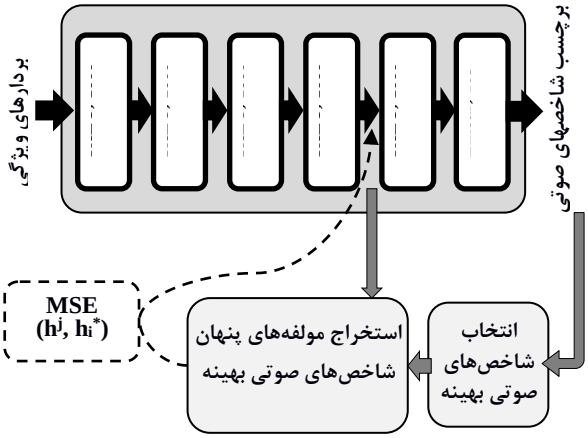
در طرح‌های جبران‌سازی تنوعات نخست و دوم از شاخص‌های بهینه برای سوق‌دادن نمونه‌های دارای تنوع به نمونه‌هایی که بهتر قابل بازشناسی هستند، استفاده شد. همان‌طور که پیش‌تر بیان شد، شاخص‌های بهینه قبل از اعمال طرح‌های جبران‌سازی با استفاده از ساختار بازشناس شاخص‌های صوتی استخراج می‌شوند و در طرح‌های نخست و دوم تغییری نمی‌کنند. نقطهٔ ضعف این روش در این است که شبکهٔ بازشناس شاخص‌های صوتی در طرح‌های نخست و دوم بهبود و اصلاح می‌شود، ولی این شاخص‌های صوتی بهینه از پیش تعیین‌شده و ثابت هستند. برای رفع مشکل یادشده در طرح چهارم از ایده مطرح‌شده در شبکه‌های عصبی عمیق جاذب‌دار (DANN) استفاده شده است.

شبکه‌های DANN ابتدا در [۳۱] برای جداسازی گویندگان پیشنهاد شد. بدین منظور در هر گامِ آموزش از مؤلفه‌های استخراج‌شده از لایهٔ پنهان انتهایی شبکه برای دادگان هر گوینده متوسط‌گیری می‌شود. این مقادیر متوسط به‌عنوان جاذب گویندگان در نظر گرفته می‌شود؛ سپس فاصله دادگان تعلیم از جاذب‌ها محاسبه و با توجه به نزدیکی و دوری از هر جاذب برچسب جدید گویندگان

[DOI: 10.61186/jsdp.19.4.173]

تخمین زده می‌شود. فاصلهٔ برچسب قدیمی و جدید برای تعلیم شبکه مورد استفاده قرار گرفته و تعلیم تا زمانی ادامه می‌یابد که مکان جاذب‌ها به‌طورتقریبی تغییر زیادی نداشته باشد. در [۲۹] هم از همین ایده برای کاربرد جداسازی آهنگ از گفتار و در [۱۳] برای جداسازی منابع و حذف انعکاس استفاده شده است..

در این پژوهش از این ایده برای استخراج شاخص‌های صوتی بهینه و شکل‌دهی مناسب مؤفه‌های لایه‌های پنهان برای حذف تنوعات ناخواسته استفاده شده است. بدین منظور در تعلیم ساختار بازشناسی شاخص‌های صوتی شکل (۶) تغییراتی اعمال شده است و ساختار با دو خطا تعلیم می‌یابد. خطای نخست خطای بازشناسی شاخص‌های صوتی است و خطای دوم، فاصلهٔ مؤلفه‌های لایهٔ پنهان شبکه از مؤلفه‌های شاخص‌های صوتی بهنیه‌ای است که در هر مرحلهٔ تعلیم از خروجی شبکه به‌دست می‌آید. واضح است که شاخص‌های صوتی بهینه در این روش از قبل ثابت نیست و در هر گام از تعلیم شبکه ممکن است نمونه‌های جدیدی به‌عنوان شاخص‌های بهینه انتخاب شود. درصورت تعلیم مناسب، شبکه می‌تواند مؤلفه‌های مشابهی را برای دادگان هم‌طبقه در لایهٔ پنهان شکل دهد. این مؤلفه‌های که متناظر با شاخص‌های بهینه هستند، مانند جاذب‌های به کار رفته در [۳۱] عمل می‌کنند.



(شکل-۱۲): طرح ۴ جبران‌سازی- شکل‌دهی مؤلفه‌های لایه‌های پنهان برای جبران‌سازی تنوعات.

(Figure-12): Scheme 4- Shaping hidden layers components to compensate variabilities.

شکل (۱۲) ساختار طرح چهارم جبران‌سازی تنوعات را نشان می‌دهد. در این شکل *H مجموعهٔ شاخص‌های صوتی بهینهٔ انتخاب‌شده در هر مرحله از تعلیم است. خطای میانگین مربعات (MSE[1]) مؤلفه‌های لایهٔ پنهان هر نمونه‌ی j ام از دادگان تعلیم (h^j) از مؤلفه‌های شاخص صوتی بهینهٔ متناظرش (h_i^*) برای شکل‌دهی مؤلفه‌های لایهٔ پنهان مورد استفاده قرار می‌گیرد.

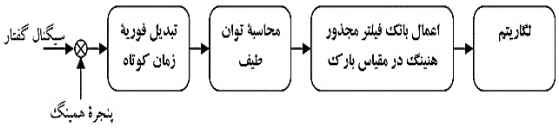
۶- پیاده‌سازی و نتایج

۶-۱- دادگان

دادگان مورد استفاده در این مقاله، دادگان گفتاری فارسی "فارس‌دات" است. این دادگان دارای ۳۰۴ گوینده و ۳۸۶ جملهٔ متفاوت است. هر گوینده بیست جمله از این جملات را در دو جلسهٔ مختلف در اتاقک دارای عایق صوتی و به‌صورت رسمی خوانده است [۱۱]. مجموعه‌دادگانِ آموزش، توسعه و آزمون مشابه [۱] به‌ترتیب ۲۲۴، ۵۰ و ۳۰ است. این دادگان به‌نحوی انتخاب شده است که به غیر از دو جملهٔ مشترک در همه گویندگان، جملات مشابه دیگری نداشته باشند.

۶-۲- استخراج ویژگی‌ها

ویژگی‌های استخراج‌شده از دادگان گفتاری در این مقاله، ویژگی‌های طیفی LHCB[2] است. در [۳] نشان داده است که این ویژگی‌ها برای آموزش به شبکه‌های عصبی نسبت به تعدادی از ویژگی‌های معمولِ دیگر کارایی بهتری دارد. نحوهٔ استخراج این ویژگی‌ها در شکل-۱۳ نشان داده شده است.



(شکل-۱۳): روش استخراج LHCB [۳].

(Figure-13): LHCB extraction method [3].

ویژگی‌های LHCB در این مقاله از قاب‌های گفتاری با طول ۲۵ میلی‌ثانیه و هم‌پوشانی ۱۰ میلی‌ثانیه از سیگنال گفتار میکروفونی با فرکانس ۴۴۱۰۰ هرتز استخراج شده است. بدین منظور، پس از محاسبهٔ طیف توان از تبدیل فوریهٔ زمان کوتاه هر قاب گفتاری، ۱۸ عدد فیلتر در مقیاس بارک[3] (در محدوده فرکانسی ۰ تا ۷۵۰۰ هرتز) اعمال شده و بر خروجی این فیلترها، لگاریتم اعمال شده است؛ سپس دلتا و دلتا-دلتا این ویژگی‌ها نیز به بردار ویژگی اضافه می‌شود. درنتیجه بردار ویژگی استخراج شده از هر قاب گفتاری، ۵۴ بُعدی است. در نهایت، بردارهای ویژگی به میانگین و انحراف معیار هنجارسازی شده است.

[1] Mean Square Error

[2] Logarithm of Hanning Critical Bands Filter Banks

[3] Bark Scale

۳-۶- معیار ارزیابی

معیار ارزیابی مدل‌ها در این مقاله، متوسط نرخ خطای بازشناسی آوایِ دادگان آزمون است. این خطا برای هر دادهٔ آزمون به‌صورت زیر محاسبه می‌شود:

$$PER_n = \frac{K - D - I - S}{K} \quad (۱)$$

در این رابطه، K تعداد کل آواهای دادهٔ آزمون n ام، D تعداد آواهای حذف‌شده، I تعداد آواهای درج‌شده، S تعداد آواهای جانشین‌شده و PER_n نرخ خطای بازشناسی آوای دادهٔ آزمون nام است.

۴-۶- نتایج بازشناسی آوا با استفاده از سامانهٔ مبتنی بر شاخص‌های صوتی

ساختار نمایش داده‌شده در (شکل-۶) با تعداد لایه‌ها، نورون‌ها و محتوای ورودی مختلف آموزش داده شد. مقصود از محتوای ورودی، تعداد قاب‌های گفتاری است که به ورودی شبکه داده می‌شود. بدین منظور، تعداد m قاب قبل و m قاب بعد هر قاب مرکزی (در مجموع $2\times m+1$ قاب) به‌عنوان ورودی به شبکه‌ها داده شده است. تعداد نورون‌های لایه‌های پنهان برابر هم انتخاب شده و به‌صورت $N \times L$ (N لایه پنهان دارای L نورون در هر لایه) بیان می‌شود.

(جدول-۳): نتایج خطای بازشناسی آوای سامانهٔ مبتنی بر شاخص‌های صوتی.

(Tabel-3): Phone error rate results of landmark based system.

	تعداد قاب ورودی	ساختار شبکه	PER
۱	13	702-500×5-103	22.29
۲	15	810-500×5-103	22.09
۳	17	918-500×5-103	22.18
۴	15	810-750×5-103	21.94
۵	15	810-1000×5-103	21.74
۶	15	810-1200×5-103	21.83
۷	15	810-1000×6-103	21.88
۸	15	810-1000×7-103	21.91

برای آموزش تمامی مدل‌های شبکه عصبی در این مقاله از کتابخانهٔ کراس[1] در پایتون[2] استفاده شده است. وزن‌های مدل‌ها از مقادیر تصادفی شروع شده و با استفاده از الگوریتم آدام[3] به‌روز می‌شود. این الگوریتم با نرخ یادگیری ۰.۰۰۱ مقداردهی اولیه شده است و درصورتی‌که در پنج دورهٔ تعلیم، تابع هزینهٔ شبکه روی دادگان توسعه بهبود خطایی نداشته باشد، نرخ خطا به میزان ۰.۱ از نرخ یادگیری قبلی کاهش می‌یابد. نرخ کاهش نمایی تکانهٔ[4] نخست و دوم این الگوریتم نیز به ترتیب برابر ۰.۹۹۹ و ۰.۹ در نظر گرفته شده است. آموزش مدل‌ها درصورت عدم بهبود خطای بازشناسی روی دادگان توسعه متوقف می‌شود. نرخ یادگیری نیز در صورت توقف آموزش روی دادگان توسعه پس از پنج تکرار، کاهش می‌یابد. معیار خطای خروجی خطای میانگین مربعات در نظر گرفته شده است.

در (جدول-۳) نتایج برخی از آزمایش‌های انجام‌شده برای تعیین ساختار بهینه آورده شده است. نتایج این جدول نشان می‌دهد که:

- بهترین میزان فریم‌های ورودی ۷ فریم قبل و بعد فریم مرکزی (مجموعاً ۱۵ فریم) است.
- با افزایش تعداد نورون‌های لایهٔ پنهان به هزار نورون، خطای بازشناسی آوا کمتر می‌شود.
- با افزایش تعداد لایه‌های پنهان تا پنج لایه، خطای بازشناسی آوا کمتر می‌شود.

با توجه به این نتایج، ساختار مشخص شده در ردیف پنجم از جدول-۳ به‌عنوان مدل بهینه انتخاب شد. روش‌های پالایش تنوعات روی این مدل اعمال شده است.

۵-۶- نتایج استفاده از روش‌های پالایش تنوعات

در بخش ۵، چهار طرح مختلف برای بازشناسی مقاوم به تنوعات شاخص‌های صوتی معرفی شد. این طرح‌ها روی بهترین مدل بازشناس شاخص‌های صوتی زیر بخش قبل اعمال شد. نتایج این آزمایش‌ها (جدول-۴) نشان می‌دهند که هر چهار طرح موجب بهبود نتایج بازشناسی می‌شوند و تا حدودی تنوعات شاخص‌های صوتی را پالایش می‌کنند. با توجه به جدول-۴ می‌توان نتایج زیر را برداشت کرد:

- افزودن حلقه بازگشتی در طرح‌های نخست و دوم مفید است.
- با توجه به نتایج طرح نخست، بازسازی کامل خروجی نیاز نیست و بهتر است از مؤلفه‌های لایهٔ پنهان استفاده شود.
- گرچه طرح سوم نیز موجب بهبود نتایج می‌شود ولی

[1] Keras
[2] Python
[3] Adam
[4] Momentum

[DOI: 10.61186/jsdp.19.4.173]

به نظر می‌رسد که در طرح دوم به دلیل استفاده از اطلاعات شاخص‌های بهینه، نتایج بهتری به‌دست آمده است.

- بهترین نتایج مربوط به طرح چهارم است که به‌طورتقریبی از تمام مزایای طرح‌های دیگر مثل استفاده از اطلاعات شاخص‌های بهینه استفاده می‌کند، ولی پارامتری نیز به وزن‌های ساختار اضافه نمی‌کند.

(جدول-۴): نتایج اعمال روش‌های پالایش تنوعات شاخص‌های صوتی

(Tabel-4): Phone error rate results of landmark based system

PER	ساختار شبکه
21.74	شبکه بازشناس بهینه
21.83	شبکه بازشناس بهینه + طرح ۱ پالایش تنوعات با خروجی از لایه‌ی آخر و بدون اتصال بازگشتی
21.57	شبکه بازشناس بهینه + طرح ۱ پالایش تنوعات با خروجی از لایه‌ی آخر و با اتصال بازگشتی
21.49	شبکه بازشناس بهینه + طرح ۱ پالایش تنوعات با خروجی از لایه‌ی پنهان و بدون اتصال بازگشتی
21.35	شبکه بازشناس بهینه + طرح ۱ پالایش تنوعات با خروجی از لایه‌ی پنهان و با اتصال بازگشتی
21.16	شبکه بازشناس بهینه + طرح ۲ پالایش تنوعات بدون اتصال بازگشتی
20.68	شبکه بازشناس بهینه + طرح ۲ پالایش تنوعات با اتصال بازگشتی
21.31	شبکه بازشناس بهینه + طرح ۳ پالایش تنوعات
20.44	شبکه بازشناس بهینه + طرح ۴ پالایش تنوعات

۶-۶- مقایسهٔ نتایج این پژوهش با دیگر سامانه‌های بازشناسی آوای فارسی

در این زیربخش برای مقایسه، نمونه‌ای از روش‌های ارائه‌شده روی دادگان فارس‌دات آورده شده است (جدول-۵). متأسفانه نتایج گزارش‌شده روی دادگان فارس دات به دلیل یکسان نبودن مجموعه‌های آموزش و آزمون برای مقالات مختلف، قابل مقایسهٔ دقیق نیستند. با این وجود با توجه به مجموعهٔ آزمون سخت‌گیرانه انتخاب‌شده در این مقاله، نتایج سامانهٔ بازشناسی آوای مبتنی بر شاخص‌های صوتی کیفیت به‌نسبه مناسبی را نشان می‌دهد. گرچه خطای بازشناسی آوای این روش در مقایسه با برخی روش‌های موجود یادشده در جدول(۵) بیشتر است، ولی نتایج به‌دست‌آمده، مفیدبودن این رویکرد متفاوت را در بازشناسی گفتار نشان می‌دهد. این نتایج از این نظر حائز اهمیت است که می‌توان با ادامه تحقیقات در این حوزه، این رویکرد را همسنگ با سایر روش‌های متداول بازشناسی گفتار رشد داد.

(جدول-۵): مقایسهٔ سامانه‌های بازشناسی آوای فارسی در دسترس

(Tabel-5): comparision of available Persian phone recognition systems

PER	روش	داده آزمون	مقاله
19.80	SGMM[1] + MMI[2]	۳۰ نفر	Babaali, 2016 [1]
19.02	MDNN[3]	۷ نفر	Ansari, 2017 [9]
26.70	HMM		
23.31	HMM	-	Firooz, 2017 [19]
21.90	CNN-RNN +CTC	۱ نفر	Alisamir, 2018 [7]
20.93	DNN- HSMM[4]	۵۰ نفر	Kermanshahi, 2019 [28]
24.80	HMM	۳۰ نفر	Veisi, 2020 [48]
16.70	DBLSTM[5] -DNN		
20.44	مبتنی بر شاخص‌های صوتی	۳۰ نفر	این مقاله

۷- جمع‌بندی

در این مقاله یک سامانهٔ بازشناسی آوای مبتنی بر استخراج مقاوم شاخص‌های صوتی ارائه شد. بدین‌منظور با بهره گرفتن از دانش زبان‌شناسی موجود و پژوهش‌های گذشته، شاخص‌های صوتی مناسب برای سیگنال گفتار زبان فارسی انتخاب شد؛ سپس ویژگی‌های طیفی مناسب به‌عنوان ورودی و برچسب شاخص‌های صوتی به‌عنوان خروجی به یک ساختار شبکهٔ عصبی جلوسو تمام متصل داده شد. بهترین نتایج با استفاده از ساختار پنج لایه پنهان با خطای بازشناسی آوای ۲۱/۷۴ به‌دست آمد؛ سپس برای افزایش مقاومت ساختار ارائه شده به تنوعات شاخص‌های صوتی، اصلاحاتی در ساختار مدل ایجاد شد. در طرح نخست با استفاده از یک شبکهٔ حذف تنوعات جداگانه، شاخص‌های صوتی به شاخص‌های صوتی متناظر بهینه‌شان نگاشت شد و خطای بازشناسی آوا به عدد ۲۱.۳۵ رسید. در طرح دوم، با آموزش هم‌زمان جبران‌ساز تنوعات و برچسب شاخص‌های صوتی، خطای بازشناسی آوا به مقدار ۲۱/۱۶ بهبود یافت. در طرح سوم از یک اتصال بازگشتی در لایهٔ پنهان شبکه برای بازسازی ورودی استفاده شده و خطای بازشناسی آوای ۲۱.۳۱ به‌دست

[1] Sub-space Guassian Mixture Model
[2] Maximum Mtutal Information
[3] Modular Deeo Neural Networks
[4] Hidden Semi-Markov Model
[5] Deep Bidirectional Long Short-Term Memory

[DOI: 10.61186/jsdp.19.4.173]

آمد؛ درنهایت در طرح چهارم با استفاده از یک ساختار مبتنی بر شبکه‌های DANN خطای بازشناسی آوا به ۲۰.۴۴ رسید. در مجموع، کارایی مناسب سامانهٔ پردازش شاخص‌های صوتی در مقایسه با سایر سامانه‌های ارائه‌شده روی سیگنال گفتارِ زبان فارسی نشان می‌دهد که این رویکرد می‌تواند به‌عنوان یک روش مؤثر برای بازشناسی گفتار مورد بررسی قرار گیرد.

۸- پیشنهادها جهت ادامهٔ کار

در این مقاله، شاخص‌های صوتی مستقل از سرعت، بیان تعریف شده‌اند و تا حد ممکن با استفاده از ساختارهای مبتنی بر شبکه‌های عصبی، سعی در حذف تنوعات ناخواسته‌ای چون تنوعات گوینده از آنها شده است؛ ازاین‌رو به نظر می‌رسد، استفاده از شاخص‌های صوتی می‌تواند موجب مقاوم‌شدن سامانهٔ بازشناسی گفتار به تنوعات ناخواسته شود. از آنجا که یکی از منابع مهم تنوعات ناخواسته نوفه‌های محیطی است، یکی از پیشنهادهای اصلی در ادامهٔ این پژوهش، مقایسه عملکرد سامانهٔ بازشناسی گفتار مبتنی بر شاخص‌های صوتی با یک سامانهٔ مبتنی بر روش‌های معمولی چون CTC در مواجهه با نوفهٔ محیطی است؛ علاوه‌بر این همان‌طور که پیش‌تر بیان شد، از محل شاخص‌های صوتی به‌عنوان نواحی پُراطلاعات گفتاری می‌توان برای تشخیص محل آواها در رشتهٔ گفتار و ایجاد هم‌ردیفی رشتهٔ ورودی و رشتهٔ خروجی استفاده کرد. با این هدف، در ادامهٔ این پژوهش سعی خواهد شد که از شاخص‌های صوتی برای رفع مشکل هم‌ردیفی در سامانه‌های بازشناسی گفتار استفاده و سامانهٔ بازشناسی گفتاری به‌صورت تلفیق روش‌های مبتنی بر CTC و شاخص‌های صوتی ارائه شود. در ادامه سه مسیر ممکن برای تلفیق اطلاعات شاخص‌های صوتی و سامانه‌های بروز بازشناسی گفتار مبتنی بر CTC پیشنهاد شده است:

- **طرح ۱:** استفاده از اطلاعات نواحی حاوی شاخص‌های صوتی برای محدودکردن جستجوی هم‌ردیفی مناسب در الگوریتم CTC و بهبود سرعت آن
- **طرح ۲:** استفاده از اطلاعات نواحی شاخص‌های صوتی به عنوان یک سیگنال وزن‌دهی جهت اعمال توجه بیشتر در محل فریم‌های حاوی اطلاعات بیشتر
- **طرح ۳:** استفاده از مؤلفه‌های استخراج‌شده از شبکهٔ بازشناس شاخص‌های صوتی به‌عنوان ویژگی کمکی در سامانهٔ بازشناسی گفتارِ مبتنی بر CTC

۹- مراجع 9_Refrence

[۱] ب. باباعلی، *پایه‌گذاری بستری نو و کارآمد در حوزهٔ بازشناسی گفتار فارسی*، مجلهٔ پردازش علائم و داده‌ها، جلد ۱۳، صفحات ۶۲-۵۱، ۱۳۹۵.

B. Babaali, A state-of-the-art and effitient framework for persian speech recognition, Signal and Data Processing, Vol. 13, pp. 51-62, 2016.

[۲] ی.ثمره، *آواشناسی زبان فارسی*، تهران، مرکز نشر دانشگاهی، ۱۳۶۴.

Y. Samareh, Persian language phonology, Tehran, university publishing center, 1985.

[۳] م. رحیمی‌نژاد و س. ع. سیدصالحی، *مقایسه و ارزیابی کارایی انواع روش‌های استخراج پارامترهای بازنمایی و هنجارسازی در بازشناسی مستقل از گویندهٔ گفتار*، نشریهٔ علمی پژوهشی امیرکبیر، ۱۳۸۲.

M. Rahiminezhad, S. A. Seyyedsalehi, Comparision and assessment of different feature extraction and normalization methods in speaker independent speech recognition, Amirkabir journal of science and research, 2000.

[۴] س. ع. سید صالحی، ا. نژادقلی، ف. توحیدخواه، *افزایش کارایی بازشناخت الگوی شبکه‌های عصبی جلوسو از طریق توسعهٔ روش‌هایی برای دوسویه کردن عملکرد آنها*، گزارش طرح مستقل پژوهشی، ۱۳۸۳.

S. A. Seyyedsalehi, I. Nejadgholi, F. Tohidkhah, Boostingt pattern recognition performance of neural networks with deleoping bidirectional methods, independent research report, 2004.

[۵] ش. کرمی، بازشناسی واج‌های گفتار پیوستهٔ فارسی به‌وسیلهٔ شبکه‌های عصبی به‌صورت مستقل از گوینده با ترکیب اطلاعات نواحی گذرا و یکنواخت واج‌ها، پایان‌نامهٔ کارشناسی ارشد مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، ۱۳۷۹.

S. Karami, Speaker independent persian phone recognition using a neural network model with a combination of steady and transition parts of phones, M.Sc. thesis, Biomedical engineering faculty, Amirkabir University, 2000.

[۶] م. یزدیان، بازشناسی گفتار پیوسته فارسی بر مبنای مدل‌سازی وقایع گسستهٔ صوتی، پایان‌نامهٔ کارشناسی ارشد مهندسی پزشکی، دانشگاه صنعتی امیرکبیر، ۱۳۸۰.

[DOI: 10.61186/jsdp.19.4.173]

M. Yazdiyan, Persian continous speech recognition based on discrete acoustic events modeling, M.Sc. thesis, Biomedical engineering faculty, Amirkabir University, 2001.

[7] S. Alisamir, S. M. Ahadi, and S. Seyedin, An end-to-end deep learning model to recognize Farsi speech from raw input, 4th Iranian Conference on Signal Processing and Intelligent Systems, pp. 1-5, 2018.

[8] N. Amini, S. A. Seyyedsalehi, Manipulation of attractors in feed-forward autoassociative neural networks for robust learning, Iranian Conference on Electrical Engineering (ICEE), 2017.

[9] Z. Ansari and S. A. Seyyedsalehi, Toward growing modular deep neural networks for continuous speech recognition, Neural Computing and Applications, pp.1177-1196, 2017.

[10] S. Babaei, , A. Geranmayeh, and S. A. Seyyedsalehi, Protein secondary structure prediction using modular reciprocal bidirectional recurrent neural networks, Computer methods and programs in biomedicine, 100(3), pp.237-247, 2010.

[11] M. Bijankhan, J. Sheikhzadegan, M. R. Roohani, FARSDAT-the speech database of Farsi spoken language, proccedings australian conference on speech science and technology, 1994.

[12] S. Borysand M. Hasegawa-Johnson, SVM-HMM landmark based speech recognition, 2009.

[13] Z. Chen, Y., Luo and N. Mesgarani, Deep attractor network for single-microphone speaker separation, In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 246-250, 2017.

[14] J. Chorowski, D. Bahdanau, K. Cho and Y. Bengio, End-to-end continuous speech recognition using attention-based recurrent NN: first results, arXiv, pp.1412.1602, 2014.

[15] G. Dahl, M. A. Ranzato, A. R. Mohamed and G. E. Hinton, Phone recognition with the mean-covariance restricted Boltzmann machine, Advances in neural information processing systems, pp. 469-477, 2010.

[16] Z. D. Doolab, S. A. Seyyedsalehi, and N. S. Dehaghani, , Nonlinear Normalization of Input Patterns to Handwritten Character Variability in Handwriting Recognition Neural Network, International Conference on Biomedical Engineering and Biotechnology, pp. 848-851, 2012.

[17] L. Dehyadegary, S. A. Seyyedsalehi and I. Nejadgholi, Nonlinear enhancement of noisy speech using continuous attractor dynamics formed in recurrent neural networks, Neurocomputing. 2011.

[18] B. Delgutte and N. Y. Kiang, Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics, The Journal of the Acoustical Society of America, pp.897-907, 1984.

[19] S. Firooz, F. Almasganj, and Y. Shekofteh, Improvement of automatic speech recognition systems via nonlinear dynamical features evaluated from the recurrence plot of speech signals, Computers & Electrical Engineering, pp. 215-226, 2017.

[20] D. Gillick, S. Wegmann and L. Gillick, Discriminative training for speech recognition is compensating for statistical dependence in the HMM framework, international conference on acoustics, speech and signal processing (ICASSP), pp. 4745-4748, 2012.

[21] A.H. Hadjahmadi, and M. M. Homayounpour, Robust feature extraction and uncertainty estimation based on attractor dynamics in cyclic deep denoising autoencoders, Neural Computing and Applications, 31(11), pp.7989-8002, 2019.

[22] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan and J. Muller, Landmark-based speech recognition, Report of the 2004 Johns Hopkins summer workshop, International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2005.

[23] D. He, B. P. Lim, X. Yang, M. Hasegawa-Johnson and D. Chen, Acoustic landmarks contain more information about the phone string than other frames for automatic speech recognition with deep neural network acoustic model, The Journal of the Acoustical Society of America, pp. 3207-3219, 2018.

[24] D. He, X. Yang, B. P. Lim, Y. Liang, M. Hasegawa-Johnson and D. Chen, When CTC training meets acoustic landmarks., International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5996-6000, 2019.

[25] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath and B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups, IEEE signal processing magazine, pp. 82-97, 2012.

[26] A. Juneja and C. Espy-Wilson, A probabilistic framework for landmark detection based on phonetic features for automatic speech recognition, journal of the acoustical society of America, pp. 1154-1168, 2008.

[27] J. Kahn, A. Lee and A. Hannun, Self-training for end-to-end speech recognition, IEEE International Conference on Acoustics,

[DOI: 10.61186/jsdp.19.4.173]

Denoising Autoencoders for Robust Phone Recognition, 29th Iranian Conference on Electrical Engineering, 2021.

[40] T. N. Sainath, Island-driven search using broad phonetic classes, automatic speech recognition & understanding, pp. 287-292, 2009.

[41] T. N. Sainath, B. Kingsbury and B. Ramabhadran, Auto-encoder bottleneck features using deep belief networks, IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4153-4156, 2012.

[42] L. San, N. Moritz, T. Hori, and J. L. Roux, Unsupervised Speaker Adaptation Using Attention-Based Speaker Memory for End-to-End ASR, International Conference on Acoustics, Speech and Signal Processing, 2020.

[43] S. A. Seyyedsalehi, A modular neural network speech recognizer based on the both acoustic steady portions and transitions, international conference of spoken language processing (ICSLP), 2000.

[44] S. Z. Seyyedsalehi, and S. A. Seyyedsalehi, Attractor analysis in associative neural networks and its application to facial image analysis, Computational Intelligence in Electrical Engineering, Vol. 9, No. 1, 2018

[45] K. N. Stevens, S. J. Keyser, and H. Kawasaki, Toward a phonetic and phonological theory of redundant features , Ph.D. thesis. MIT, camberidge, 1986.

[46] K. N. Stevens, From acoustic cues to segments, features and words, international conference on Spoken Language Processing (ICSLP), pp. A1–A8, 2000.

[47] J. Vaněk, J. Michálek and J. Psutka, Recurrent DNNs and Its Ensembles on the TIMIT Phone Recognition Task, International Conference on Speech and Computer, pp. 728-736, 2018.

[48] H. Veisi, and A. H. Mani, Persian speech recognition using deep learning, International Journal of Speech Technology, 23(4), pp. 893-905, 2020.

[49] P. Vincent, H. Larochelle, Y. Bengio and P. A. Manzagol, Extracting and composing robust features with denoising autoencoders, Proceedings of the 25th international conference on Machine learning, pp. 1096-1103), 2008.

[50] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P. A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, Journal of machine learning research, pp. 3371-3408, 2010.

[51] M. Zapotoczny, P. Pietrzak, A. Lancucki and J. Chorowski, 2019. Lattice generation in Speech and Signal Processing (ICASSP), pp. 7084-7088, 2020.

[28] M. A. Kermanshahi and M. M. Homayounpour, Improving Phoneme Sequence Recognition using Phoneme Duration Information in DNN-HSMM, Journal of AI and Data Mining, pp.137-147, 2019.

[29] R., Kumar, Y. Luo, and N. Mesgarani, Music Source Activity Detection and Separation Using Deep Attractor Network, In INTERSPEECH, pp. 347-351, 2018.

[30] J. W. Lee, J. Y. Choi and H. G. Kang, Classifcation of stop place in consonant-vowel contexts using feature extrapolation of acoustic-phonetic features in telephone speech, The Journal of the Acoustical Society of America, Vol. 131, 2012.

[31] Y. Luo, Z. Chen and N. Mesgarani, Speaker-independent speech separation with deep attractor network, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(4), 787-796, 2018.

[32] M. Meister and M. J. Berry, The neural code of the retina, Neuron, pp.435-450, 1999.

[33] N. Morgan, J. Cohen, S. H. Krishnan, S. Changand S. Wegmann, Final Report: OUCH Project (Outing Unfortunate Characteristics of HMMs), 2013.

[34] T.S. Nguyen, S. Stüker, J. Niehues, and A. Waibel, Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation, IEEE International Conference on Acoustics, Speech and Signal Processing, 2020.

[35] C. Niu, J. Zhang, X. Yang and Y. Xie, A study on landmark detection based on CTC and its application to pronunciation error detection, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 636-640, 2017.

[36] S. Parveen, and P. Green, Speech enhancement with missing data techniques using recurrent neural networks, IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1, pp. I-733, 2004.

[37] M. Ravanelli, T. Parcollet and Y. Bengio, The pytorch-kaldi speech recognition toolkit, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6465-6469, 2019.

[38] S. Reza, S. A. Seyyedsalehi, S. Z. Seyyedsalehi, A Persian Language Phone Recognition Based on Robust Extraction of Acoustic Landmarks, 27th Iranian Conference on Biomedical Engineering, 2020.

[39] S. Reza, S. A. Seyyedsalehi, S. Z. Seyyedsalehi, Attractor Manipulation in

[DOI: 10.61186/jsdp.19.4.173]

attention-based speech recognition models, pp.2225-2229, 2019.

[52] T. Yoshimura, T. Hayashi, K. Takeda and S. Watanabe, End-to-end automatic speech recognition integrated with CTC-based voice activity detection, arXiv, 2020.

[53] N. Zeghidour, N. Usunier, I. Kokkinos, T. Schaiz, G. Synnaeve and E. Dupoux, Learning filterbanks from raw speech for phone recognition, International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5509-5513, 2018.

شقایق رضا: تحصیلات خود را در مقطع کارشناسی در رشته مهندسی پزشکی – بیوالکتریک در دانشگاه صنعتی امیرکبیر در سال ۱۳۸۵ و کارشناسی ارشد را در همان رشته در دانشگاه صنعتی امیرکبیر در سال ۱۳۸۷ به پایان رساند. ایشان هم‌اکنون دانشجوی مقطع دکترای بیوالکتریک در دانشگاه صنعتی امیرکبیر است. از موضوعات مورد علاقه ایشان پردازش گفتار، شبکه‌های عصبی مصنوعی و یادگیری عمیق است. نشانی رایانامه ایشان عبارت است از: sh.reza@aut.ac.ir

سیّد علی سیّدصالحی: مدرک کارشناسی خود را در رشتهٔ مهندسی برق از دانشگاه صنعتی شریف در سال ۱۳۶۱، کارشناسی ارشد را در همان رشته از دانشگاه صنعتی امیرکبیر در سال ۱۳۶۷ و دکترای خود را در رشتهٔ مهندسی پزشکی- بیوالکتریک از دانشگاه تربیت مدرس در سال ۱۳۷۴ دریافت کرده‌اند. ایشان در حال حاضر دانشیار دانشکدهٔ مهندسی پزشکی دانشگاه صنعتی امیرکبیر هستند. زمینه‌های پژوهشی مورد علاقهٔ ایشان پردازش و بازشناسی گفتار، شبکه‌های عصبی مصنوعی و زیستی، مدل‌سازی عملکرد مغز و پردازش خطی و غیرخطی سیگنال است.
نشانی رایانامهٔ ایشان عبارت است از: ssalehi@aut.ac.ir

سیّده زهره سیّدصالحی: مدرک کارشناسی خود را در رشتهٔ مهندسی پزشکی – بیوالکتریک از دانشگاه امیرکبیر در سال ۱۳۸۳، کارشناسی ارشد را در همان رشته از دانشکدهٔ فنی دانشگاه شاهد در سال ۱۳۸۶ و دکترای مهندسی پزشکی را در سال ۱۳۹۲ از دانشگاه صنعتی امیرکبیر با تمرکز بر فضای یادگیری عمیق به پایان رسانده است. ایشان همچنین از سال ۱۳۹۴ به‌مدت دو سال در دانشکده مهندسی کامپیوتر دانشگاه صنعتی شریف پژوهش‌گر پسادکترا بوده‌اند. زمینه‌های پژوهشی مورد علاقهٔ ایشان یادگیری عمیق و شبکه‌های عصبی مصنوعی است.
نشانی رایانامه ایشان عبارت است از:
z.seyyedsalehi@aut.ac.ir

[DOI: 10.61186/jsdp.19.4.173]