



تشخیص عبارت‌های گفتاری برای اخبار فارسی صداوي‌سیما جمهوری اسلامی ایران

هادی ویسی^۱، سید اکبر قریشی^۲ و اعظم باستان فرد^۳

^۱ دانشکده علوم و فنون نوین، دانشگاه تهران، تهران، ایران

^۲ دانشکده فنی و مهندسی رسانه، دانشگاه صداوسیما، تهران، ایران

^۳ دانشگاه آزاد اسلامی واحد کرج، کرج، ایران

چکیده

هدف از تشخیص عبارت‌های گفتاری یا جستجوی کلیدواژه، تشخیص و جستجوی مجموعه‌ای از کلیدواژه‌ها در مجموعه‌ای از اسناد گفتاری (مانند سخنرانی‌ها، جلسه‌ها) است. در این پژوهش تشخیص عبارت‌های گفتاری فارسی برپایه سامانه‌های بازناسنی گفتار با کاربرد در بازیابی اطلاعات در بایگانی‌های گفتاری و ویدئویی سازمان صدا و سیما طراحی و پیاده‌سازی شده است. برای این کار، ابتدا اسناد گفتاری به متن، بازناسنی، سیس بر روی این متون جستجو انجام می‌شود. برای آموزش سامانه بازناسنی گفتار فارسی، دادگان فارس دات بزرگ به کار رفته است. این سامانه به نرخ خطای واژه ۲/۷۱ درصد بر روی همین دادگان و ۲۸/۲۳ درصد بر روی دادگان اخبار فارسی با استفاده از مدل زیر فضای مخلوط گوسی (SGMM) رسید. برای تشخیص عبارت‌های گفتاری از دوش پایه واژگان نماینده استفاده شده و با استفاده از شبکه حافظه کوتاه-مدت ماندگار و دسته‌بندی زمانی پیوندگرا (LSTM-CTC) روشی برای بهبود تشخیص واژگان خارج از واژگان (OOV) پیشنهاد شده است. کارایی سامانه تشخیص عبارات با روش واژه‌های نماینده بر روی دادگان فارس دات بزرگ بر طبق معیار ارزش وزنی واقعی عبارت (ATWV) برابر با ۹۲۰۶/۰ برای کلیدواژه‌های داخل واژگان و برابر با ۰/۰ برای کلیدواژه‌های خارج از واژگان رسید که این نرخ برای واژگان OOV با استفاده از روش LSTM-CTC با حدود پنجاه درصد بهبود به مقدار ۳۰۵۸/۰ رسید؛ همچنین، در تشخیص عبارت‌های گفتاری بر روی دادگان اخبار فارسی، برابر با ۸۰۰۸/۰ حاصل شد.

واژگان کلیدی: تشخیص عبارت‌های گفتاری فارسی، جستجوی کلیدواژه، بازناسنی گفتار، سازمان صداوسیما، کلیدی

Spoken Term Detection for Persian News of Islamic Republic of Iran Broadcasting

Hadi Veisi^{*1}, Sayed Akbar Ghoreishi² & Azam Bastanfard³

¹Faculty of New Sciences and Technologies, University of Tehran, Tehran, Iran

²Department of Media Engineering, IRI Broadcast University, Tehran, Iran

³ Karaj Islamic Azad University, Karaj, Iran

Abstract

Islamic Republic of Iran Broadcasting (IRIB) as one of the biggest broadcasting organizations, produces thousands of hours of media content daily. Accordingly, the IRIB's archive is one of the richest archives in Iran containing a huge amount of multimedia data. Monitoring this massive volume of data, and brows and retrieval of this archive is one of the key issues for this broadcasting. The aim of this research is to design a content retrieval engine for the IRIB's media and production using spoken term detection (STD) or keyword spotting. The goal of an STD system is to search for a set of keywords in a set of speech documents. One of the methods for STD is using a speech recognition system in which speech is recognized and converted into text and then, the text is searched for the keywords. Variety of speech documents and the limitation of speech recognition vocabulary are two challenges of this approach. Large vocabulary

* Corresponding author

*نویسنده عهده‌دار مکاتبات

continuous speech recognition systems (LVCSR) usually have limited but large vocabulary and these systems can't recognize out of vocabulary (OOV) words. Therefore, LVCSR-based STD systems suffer OOV problem and can't spotting the OOV keywords. Methods such as the use of sub-word units (e.g., phonemes or syllables) and proxy words have been introduced to overcome the vocabulary limitation and to deal with the out of vocabulary (OOV) keywords.

This paper proposes a Persian (Farsi) STD system based on speech recognition and uses the proxy words method to deal with OOV keywords. To improve the performance of this method, we have used Long Short-Term Memory-Connectionist Temporal Classification (LSTM-CTC) network.

In our experiments, we have designed and implemented a large vocabulary continuous speech recognition systems for Farsi language. Large FarsDat dataset is used to train the speech recognition system. FarsDat contains 80 hours voices from 100 speakers. Kaldi toolkit is used to implement speech recognition system. Since limited dataset, Subspace Gaussian Mixture Models (SGMM) is used to train acoustic model of the speech recognition. Acoustic model is trained based context tri-phones and language model is probability tri-gram words model. Word Error Rate (WER) of Speech recognition system is 2.71% on FARSDAT test set and also 28.23% on the Persian news collected from IRIB data.

Term detection is designed based on weighted finite-state transducers (WFST). In this method, first a speech document is converted to a lattice by the speech recognizer (the lattice contains the full probability of speech recognition system instead of the most probable one), and then the lattice is converted to WFST. This WFST contains the full probability of words that speech recognition computed. Then, text retrieval is used to index and search over the WFST output. The proxy words method is used to deal with OOV. In this method, OOV words are represented by similarly pronunciation in-vocabulary words. To improve the performance of the proxy words methods, an LSTM-CTC network is proposed. This LSTM-CTC is trained based on characters of words separately (not a continuous sentence). This LSTM-CTC recomputed the probabilities and re-verified proxy outputs. It improves proxy words methods due to the fact that proxy words method suffers false alarms. Since LSTM-CTC is an end-to-end network and is trained based on the characters, it doesn't need a phonetic lexicon and can support OOV words. As the LSTM-CTC is trained based on the separate words, it reduces the weight of the language model and focuses on acoustic model weight.

The proposed STD achieve 0.9206 based Actual Term Weighted Value (ATWV) for in vocabulary keywords and for OOV keywords ATWV is 0.2 using proxy word method. Applying the proposed LSTM-CTC improves the ATWV rate to 0.3058. On Persian news dataset, the proposed method receives ATWV of 0.8008.

Keywords: Persian Spoken Term Detection, IRIB, Persian News, Keyword Spotting, Speech Recognition, Kaldi

اظهار می‌کند و نتایج مرتبط با نیاز کاربر در قالب چند رسانه‌ای نمایش داده می‌شود [2]. ابزار و روش‌های پایه جستجو در محتواهای گفتاری را تشخیص عبارت‌های گفتاری (STD)⁶ می‌نامند. طبق تعریفی که مؤسسه ملی فناوری و استانداردهای آمریکا (NIST)⁷ از تشخیص عبارت‌های گفتاری ارائه کرده است، هدف از تشخیص عبارت‌های گفتاری، تشخیص سریع وجود یک عبارت-یک دنباله از واژه‌های گفتاری متولی- در یک پیکره بزرگ صوتی از داده‌های گفتاری ناهمگون است [3].

با توجه به گسترش روزافون محتوای چند رسانه‌ای در بسترهای مختلف مانند پخش همگانی و اینترنت، جستجو برتر آن‌ها اهمیت بیشتری پیدا می‌کنند، جستجوهایی مانند جستجو بر مبنای بافت صدا، تصویر و ویدئو. تشخیص عبارت‌های گفتاری این امکان را فراهم می‌کند که هر سند گفتاری، حاوی هر عبارت از نمایش متنی آن، بازیابی شود؛ درواقع تشخیص عبارت‌های گفتاری مرحله مقدماتی برای

۱- مقدمه

۱-۱- تعریف مسئله

امروزه با گسترش فضای مجازی و شبکه‌های تلویزیونی با حجم انبوهی از اطلاعات و محتوای چند رسانه‌ای روبرو هستیم و این حجم به سرعت رو به رشد است؛ همچنین حجم انبوهی از اطلاعات در بایگانی ذخیره شده‌اند. با توجه به ارزش فراوان اطلاعاتی این داده‌ها، نمایه کردن^۱، بازیابی^۲، جستجو^۳ و مرور^۴ این اطلاعات، امروزه بسیار مورد توجه قرار گرفته است [1]. هرچند امروزه بازیابی اطلاعات متنی مبتنی بر کلیدواژه موضوعی عمومی و حل شده به نظر می‌رسد؛ اما جستجو در محتواهای غیرمتنی مانند گفتار یکی از موضوعات مهم پژوهشی است. هدف جستجو بر روی گفتار، بازیابی محتوای گفتاری از منابع صوتی، منطبق با نیاز است؛ کاربر نیاز خود را به صورت یک درخواست (پرسش)^۵

¹ index

² retrieve

³ search

⁴ browse

⁵ query

⁶ Spoken Term Detection (STD)

⁷ National Institute of Standards and Technology (NIST)



برای نیل به هدف موردنظر، در این مقاله بر روی اخبار فارسی صداوسیما تمرکز و از سامانه بازشناسی گفتار استفاده شده است تا گفتار به متن بازشناسی شود؛ سپس با استفاده از الگوریتم مبدل عامل^۵ [10] که یک الگوریتم برای جستجو بر روی متن است، کمک گرفته شده است تا بر روی متن بازشناسی شده، جستجو انجام پذیرد. در این روش فقط واژگان تعریف شده از قبل در سامانه بازشناسی گفتار قابل تشخیص هستند. برای جستجوی واژگان خارج از واژگان (OOV)^۶، سامانه بازشناسی گفتار واژه‌های نماینده به کار گرفته شده‌اند و برای بهبود این روش از یک شبکه حافظه کوتاه-مدت ماندگار با دسته‌بندی زمانی پیوندگرا (LSTM-CTC)^۷ استفاده شده است. در این روش نتایج حاصل از واژه‌های نماینده دوباره بازشناسی و امتیاز اطمینان آن‌ها دوباره تخمین زده می‌شود.

در ادامه مقاله و در بخش دوم، مختصراً از پژوهش‌های پیشین بر روی موضوع تشخیص عبارت‌های گفتاری بحث و سپس در بخش سوم تعریف دقیق‌تری از تشخیص عبارت‌های گفتاری عنوان می‌شود و الگوریتم‌های شاخص‌گذاری، جستجو به صورت مختصر مورد بررسی قرار می‌گیرد. در بخش چهارم روش پیشنهادی برای بازیابی اطلاعات گفتاری توضیح داده و معیارهای ارزیابی و نتایج در بخش پنجم بیان می‌شود. درنهایت بخش ششم به جمع‌بندی و نتیجه‌گیری پژوهش می‌پردازد.

۲- مروری بر پژوهش‌های پیشین

جستجو در گفتار به طور تقریبی همزمان با پژوهش‌های بازشناسی گفتار آغاز شد [11]؛ نزدیک به چهل سال پژوهش و توسعه بر روی این عنوان انجام شده و روش‌های زیادی برای آن ارائه شده است که در [1]، [2]، [5] تاریخچه بحث را می‌توانید مطالعه کنید. در پایین در سه گروه خلاصه این روش‌ها آورده شده است.

۱- روش‌های پرسش با مثال^۸

روش‌های پرسش با مثال از جمله نخستین تلاش‌ها برای جستجوی کلیدواژه‌ها هستند [13] و اسم پرسش با مثال این روش را توصیف می‌کند: نمونه‌هایی از کلیدواژه‌ها

بازیابی استناد گفتاری است [4]؛ لذا کاربردهای اصلی STD را می‌توان نمایه کردن صوت^۹ و داده‌کاوی گفتار^{۱۰} دانست.

چالش‌های اصلی STD را می‌توان به دو دسته کلی تقسیم کرد: یکی گستردگی و تنوع سندهای گفتاری است که می‌توانند به هر صورت مانند گفتارهای پیوسته، گفتارهای گسسته، گفتارهای تمیز و یا همراه با نویه باشند؛ یکی دیگر از چالش‌ها سرعت و دقت راه‌کارهای ارائه شده است. این راه‌کارها به نسبت راه‌کارهای بازیابی متن هنوز قدرتمند و دقیق نشده‌اند [1]؛ رویکردهای جستجو در گفتار را می‌توان در دو دسته کلی تقسیم‌بندی کرد:

- روش‌های تطبیق مستقیم آوایی مانند پیچش زمانی بoya^{۱۱} که در آن‌ها سعی بر تطبیق مستقیم ویژگی‌های آوایی بین گفتار و کلیدواژه (پرسش یا درخواست) است.

- استفاده از سامانه‌های خودکار بازشناسی گفتار (ASR)^{۱۲} که در این رویکرد ابتدا گفتار توسط سامانه بازشناسی خودکار گفتار، به متن تبدیل می‌شود؛ سپس با به کارگیری روش‌های بازیابی متن، بر روی متن خروجی ASR جستجو انجام می‌شود.

مطالعات گسترده‌ای برای بازیابی گفتار در روی زبان‌ها با منابع وسیع مانند زبان انگلیسی [3]، [5] انجام گرفته است و در سال‌های اخیر پژوهش‌ها بر روی زبان‌ها با منابع محدود مانند پشتون و ویتنامی نیز وجود داشته‌اند و به نتایج مطلوب و قابل قبول رسیده‌اند [6]-[8]. بالین وجود، پژوهش‌ها بر روی زبان فارسی محدود و همراه با دادگان محدود بوده‌اند [9]؛ با توجه به اهمیت این موضوع در زبان فارسی و بهخصوص برای جستجو در منابع دادگانی بزرگ مانند آرشیو سازمان صداوسیما، در این پژوهش بر روی بازیابی اطلاعات از اخبار فارسی صداوسیما پرداخته شده است. بایگانی سازمان صداوسیما یکی از منابع مهم دادگانی گفتاری در زبان فارسی برای جستجو و بازیابی اطلاعات است که با توجه به حجم انبوهای تولیدی روزانه این سازمان، هر روز به مقدار داده‌های این آرشیو افزوده می‌شود و بازیابی و استخراج اطلاعات از آن یکی از نیازهای سازمان صداوسیما است که در آینده بر اهمیت آن نیز افزوده خواهد شد. کاربردهایی از جمله جستجو در بین بایگانی‌ها و دسترسی به محتوای بایگانی، تحلیل محتوای برنامه‌های مختلف، نظارت و ارزیابی بر محتوا را می‌توان برای این موضوع در نظر گرفت.

^۱ audio indexing

^۲ speech data mining

^۳ Dynamic Time Warping(DTW)

^۴ Automatic Speech Recognition (ASR)



نمایه‌گذاری صدا و داده‌کاوی گفتار به‌طور گسترده استفاده شده‌اند. در این رویکرد، گفته‌ها توسط سامانه بازشناسی گفتار به‌واژه‌ها بازشناسی، برای جستجوی کارآمد نمایه و بر روی نمایه‌ها جستجو انجام می‌شود [5]. درواقع در این روش، سامانه LVCSR و سامانه جستجوگر باهم متوالی می‌شوند و سامانه جستجوگر بر روی خروجی بازشناسی‌شده جستجو انجام می‌دهد. در سال‌های اخیر، با این رویکرد بر روی زبان‌هایی با منابع محدود تمرکز شده است. زبان‌هایی با منابع آموزشی محدود، با حداقل چهل ساعت سند آموزشی گفتاری مانند پشتون و گرجی، هستند [6]-[8]. در پروژه IARPA babel [29] نیز مجموعه‌ای از زبان‌های محدود جمع‌آوری شده است.

به کارگیری بازشناسی گفتار در تشخیص عبارت گفتاری اگرچه بسیار رایج است، اما چالش‌های مختلفی دارد. نخستین چالش در این روش، خروجی غیرقطعی و دارای خطای سامانه بازشناسی گفتار است. نخستین بهترین⁶ فرضیه در سامانه‌های بازشناسی گفتار به‌طورعموم شامل خطاهایی هستند که عملکرد تشخیص را دچار مشکل می‌کنند. برای عملکرد بهتر، به‌طورعموم شبکه درهم‌ریختگی⁷ یا موقعیت خاص پسین شبکه⁸ [30] و زمان تدریجی شبکه⁹ [31] یا شبکه¹⁰ [32] به‌جای نخستین بهترین فرضیه‌ها برای شاخص گذاری تولید می‌شوند. یکی از روش‌ها برای غلبه بر این مشکل، ندیدن دو سامانه بازشناسی و شاخص‌گذار به‌صورت مجزا است؛ درواقع باید ترکیب متوالی این دو سامانه را به‌عنوان یک سامانه در نظر گرفت. در [1] پنج روش کلی را برای دیدن یک‌جا این سامانه‌ها معرفی کرده و موردبحث قرار داده شده است. در [5] خلاصه‌ای از کنفرانس‌هایی با موضوع بازیابی اسناد گفتاری در اخبار پخش همگانی¹¹ در سال ۲۰۰۰ آمده است که رویکرد این پژوهش‌ها به کارگیری روش‌ها و سامانه‌های بازشناسی گفتار و بازیابی متن به‌صورت متوالی برای حل این مسأله بوده است. در این مقاله عنوان شده که عملکردی مشابه نتایج انسانی در بازیابی اسناد گفتاری به‌دست آمده است که برای رسیدن به این نتایج سامانه بازشناسی گفتار نرخ خطای واژه (WER)¹² بین ۱۵ الی ۲۰ درصد باید داشته باشد، که از دقت رونوشت‌های دستی زیاد فاصله ندارد.

⁶1-best⁷Confusion network⁸Position specific posterior lattices⁹Time-quantized lattices¹⁰Lattice¹¹Broadcast¹²Word Error Rate (WER)

به‌طورعموم در قالب صدا، برای تشخیص کلیدواژه مورداستفاده قرار می‌گیرند. این روش‌ها به‌طورعمومی شامل دو مرحله هستند: مرحله بازنمایی الگو که نمونه‌های صوتی کلیدواژه‌ها به‌صورت الگوها در قالب خاصی (ویژگی‌های بسیان، مشبک‌ها) نمایش داده می‌شوند و مرحله تطبیق الگو، که در آن الگوها با گفته‌های¹ گفتاری هدف مقایسه می‌شوند. در طی دهه‌های گذشته، تمرکز پژوهش در وهله نخست بر روش‌های جدید بازنمایی الگو بوده است [14]-[19]؛ در حالی که اغلب این روش‌ها از برخی از انواع پیچش زمانی پویا² [20] برای تطبیق الگو [21] استفاده می‌کنند. در روش‌های اخیر از روش‌های مبتنی بر شبکه‌های عصبی برای آموزش الگو و تطبیق نمونه‌های صوتی استفاده شده است [22]-[23].

۲-۲- روش‌های کلیدواژه/فیلتر

در روش کلیدواژه/فیلتر که گاهی اوقات به‌عنوان تشخیص آوایی کلیدواژه شناخته می‌شود [12]، کلیدواژه و غیرکلیدواژه (فیلتر) را به‌صورت موازی مدل می‌کنند. در مرحله تشخیص، گفته‌های هدف هم با مدل کلیدواژه و هم با مدل فیلتر هم‌تاز و بر اساس هزینه هم‌تازی در مورد آنها تصمیم‌گیری می‌شود. در [15]، [16] از مدل‌های مخفی مارکوف (HMM)³ برای مدل‌کردن کلیدواژه‌ها و فیلترها استفاده شده است و با جستجو درون یک گراف رمزگشایشده که در آن کلیدواژه‌ها و فیلترها هم‌زمان ظاهر می‌شوند، تشخیص انجام می‌شود. پژوهش‌های اخیر در این زمینه کمابیش این چارچوب را همراه با تمرکز بر مدل‌کردن واژگان فیلتر [26] و روش امتیازدهی پیشرفته [27]، دنبال می‌کنند. روش‌های آموزش تمايزی⁴ [28] نیز در این زمینه مورد بررسی قرار گرفته‌اند تا به‌طور مستقیم کارایی کلیدواژه‌ها را بهینه کنند.

۲-۳- روش‌های بازشناسی گفتار پیوسته با واژگان بزرگ

روش‌های بازشناسی گفتار پیوسته با واژگان بزرگ⁵ در همین اواخر برای کاربردهایی مانند

¹Utterance²Dynamic Time Warping (DTW)³Hidden Markov Model (HMM)⁴Discriminative⁵Large vocabulary continuous speech recognition (LVCSR)



با اختصار برای مقایسه نقاط قوت و ضعف روش‌های اشاره‌شده در بالا، به موارد زیر می‌شود اشاره کرد:
در روش‌های پرسش با مثال و کلیدواژه/فیلتر نیاز به مجموعه‌ای از صوت‌ها حاوی واژه جستجو داریم، مجموعه کلیدواژه‌های مورد جستجو محدود هستند و باید از قبل مشخص باشند؛ ولی این روش‌ها سرعت بیشتر و محاسبات کمتر نسبت استفاده از سامانه بازناسی گفتار دارند؛ روش کلیدواژه/فیلتر دقت بیشتری نسبت به روش پرسش با مثال دارد.

در روش بازناسی گفتار نیاز به یک سامانه بازناسی گفتار است؛ این سامانه‌ها به طور معمول پیچیده و دارای محاسبات زیادی هستند؛ در ضمن آموزش آن‌ها زمان بر و پرهزینه است؛ ولی با استفاده از سامانه بازناسی گفتار مجموعه کلیدواژه‌ها دامنه بیشتری را دارند (روش‌های پرسش با مثال و کلیدواژه/فیلتر حداکثر چند ده تا چند صد کلیدواژه را در بردارند؛ ولی سامانه‌های بازناسی گفتار قابلیت بازناسی تا حدود چند میلیون واژه را دارند)؛ هرچند واژه‌های خارج واژگان قابل شناسایی برای سامانه بازناسی گفتار نیستند، ولی روش‌هایی برای مقابله با این مسئله ارائه شده است. در روش‌های پرسش با مثال و کلیدواژه/فیلتر برای اضافه کردن یک کلیدواژه جدید نیاز به بازنطراحی ولی در سامانه‌های بازناسی گفتار با طراحی اولیه قابلیت جستجوی بسیاری از کلیدواژه‌ها است. در این پژوهش با توجه به بازبودن دامنه کلیدواژه‌ها از سامانه بازناسی گفتار استفاده شده است. در ضمن سامانه بازناسی گفتار اطلاعات بیشتری از محتوای صوت تولید می‌کند (متن صوت را تولید می‌کند) که برای نمایش دادن به کاربر و طبقه‌بندی اطلاعات مفید است.

۴-۲- کارهای انجام گرفته روی زبان فارسی

در زبان فارسی پژوهش‌های کمی بر روی تشخیص عبارات گفتاری انجام شده است. بر طبق جستجوهای انجام گرفته توسط پژوهش‌گران این مقاله، در [9]-[34]-[38] از رویکرد به کارگیری بازناسی گفتار برای تشخیص عبارت‌های گفتاری در زبان فارسی استفاده شده است. مقاله‌ها و پژوهش‌های دیگری با رویکردهای متفاوت و عدم به کارگیری بازناسی گفتار انجام شده‌اند [42].

یکی دیگر از معایب روش‌های تشخیص کلیدواژه بر پایه‌ی بازناسی گفتار، وجود کلیدواژه‌های خارج از واژگان (OOV) است. واژگان و لغتنامه سامانه‌های بازناسی گفتار به طور معمول از قبل تعریف شده هستند و اگر یک کلیدواژه خارج از واژگان سامانه باشد، راهی برای اینکه سامانه بتواند چیزی در مورد آن کلیدواژه ارائه کند، وجود ندارد. برای غلبه بر این چالش، یکی از روش‌ها، استفاده از زیرواژه^۱ است. واحدهای زیرواژه‌ها می‌توانند واج، هجا و یا قطعه‌واژه باشند. مشبک‌های زیرواژه می‌توانند با استفاده از واژگان زیرواژه همراه با مدل‌های زبانی زیرواژه‌ای تولید شوند. همچنین می‌توان مشبک‌ها در سطح واژه را به طور مستقیم به مشبک‌های زیرواژه‌ها تبدیل شوند. بعد از این که مشبک‌های زیرواژه به وجود آمدند می‌توان بر روی دنباله زیرواژه‌ها جستجو انجام شود [25]-[26]. در [2] روش‌های ارائه شده مبتنی بر زیرواژه‌ها و انواع آن‌ها در زبان‌های مختلف به طور مفصل‌تری بحث شده‌اند. یکی دیگر از روش‌های پیشنهادشده برای حل مشکل کلیدواژه‌های OOV، روش واژه‌های نماینده^۲ [31] است. واژه‌های نماینده، واژه‌هایی هستند که شباهت آوایی با کلیدواژه OOV دارند. این روش بر پایه مبدل‌های حالت محدود وزنی است و در آن با استفاده از واژه‌های واژگان سامانه بازناسی گفتار واژه‌هایی با شباهت آوایی مناسب با کلیدواژه OOV ساخته می‌شود.

ساختمار اسناد گفتاری یکی دیگر از چالش‌ها است. این چالش در اسناد گفتاری بلند مانند پردازش اسناد مربوط به بایگانی، بیشتر مربوط می‌شود. در این حالت ممکن است یک سند گفتاری شامل موضوعات مختلف، بافت و زمینه‌های مختلف باشد.

چالش دیگر رابط کاربری سامانه است. در سامانه‌های جستجوی متنی بر متن قسمتی از متن برای کاربر نوشته می‌شود و کاربر می‌تواند سریع نتایج را بررسی کند؛ ولی در این سامانه‌ها رابط کاربری باید چگونه باشد؟ و کاربر چه طور می‌تواند نتایج را بررسی کند؟ در [1]-[2]-[27] در مورد این چالش بیشتر بحث شده است.

چالش‌های دیگری مانند سرعت جستجو، دقت جستجو و حجم خروجی شاخص گذاری شده نیز می‌توانند مطرح شوند [36].

¹ Subword² Proxy words

(جدول-۱): خلاصه‌ای از پژوهش‌های پیشین بر روی زبان‌های مختلف.

(Table-1): Summery of related works on different languages.

زبان	دادگان	روش	ATWV
انگلیسی [3]	گفتار تلفنی محاوره‌ای	ASR	0.8335
	اخبار پخش همگانی	ASR	0.8485
	اتاق‌های میزگرد گفتگو	ASR	0.2553
اسپانیایی [4]	MAVIR database	ASR	0.6404
	EPIC	ASR	0.9362
منابع محدود	۲۳ ساعت از زبان‌های منابع محدود مانند آلبانی، اسلواکی، چک، رومانی، بهصورت زبان محاوره‌ای و اخبار همگانی	پرسش-بازنمونه DTW [7]	0.58
	IARPA BABEL [29]	ASR-kaldi [6]	0.705
فارسی	فارس‌دات بزرگ [44]	ASR-BBN[8]	0.739
	فارس‌دات کوچک [43]	ASR-proxy [9]	0.85
	فارس‌دات تلفنی [45] و دادگان تلفنی بانک	phone search & WFST [38]	0.8

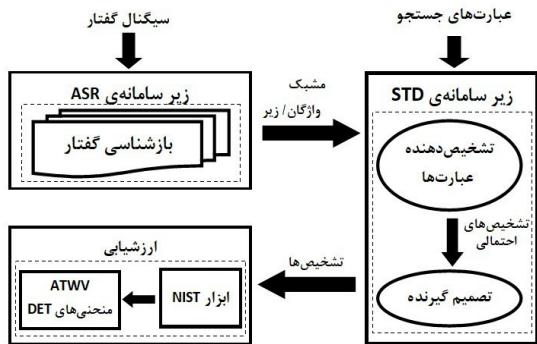
این دو نمایه جستجو می‌شوند و از روش واژه‌های نماینده برای بهبود کلیدواژه‌های OOV استفاده شده است. در این پژوهش از دادگان فارس‌دات بزرگ [44] استفاده شده و به نرخ خطای واژه ۰/۰۰ درصد بر روی فارس‌دات بزرگ رسیده و برای تشخیص عبارت‌های گفتاری با معیار ارزش بیشینه وزنی عبارت (MTWV)^۴ عدد ۸۳/۰ فقط با جستجو بر روی نمایه واژه، عدد ۳۹/۰ فقط با جستجو بر روی نمایه‌ی واژی و عدد ۸۵/۰ با ترکیب این دو جستجو حاصل شده است. در [38] عباسیان روش‌های مبتنی بر رشته واژ و جستجو مبتنی بر مشبک واژ را بر روی زبان فارسی پیاده‌سازی کرده است. در جستجوی مبتنی بر رشته واژ ابتدا گفتار به یک رشته واژ بازنیستی می‌شود؛ سپس با استفاده از فاصله لونشتاین^۵، فاصله رشته واژ بازنیستی شده با رشته واژ کلیدواژه مقایسه می‌شود. در جستجوی مبتنی بر مشبک واژ، بعد از بازنیستی گفتار به مشبک واژ، بر روی آن جستجو انجام می‌شود. این پژوهش بر روی دادگان فارس‌دات تلفنی [45] و دادگان تلفنی بانکی انجام شده و به دقت ۸۳ درصدی برای بازنیستی واژ با روش شبکه حافظه کوتاه‌مدت ماندگار دوسویه^۶ رسیده است. تعداد کلیدواژه‌های این پژوهش چهارده کلیدواژه پر تکرار بوده و در بهترین حالت به ۸/۰ برای معیار ATWV رسیده است.

در [9] گمار پژوهشی را بر روی تشخیص عبارت‌های گفتاری بر پایه بازنیستی گفتار و شرایط محدود، انجام داده است. در این پژوهش با آموزش یک شبکه گلوگاهی عصبی^۷ دقت بازنیستی واژ را بالا برده و عملکرد تشخیص عبارت‌های گفتاری را بهبود داده است. برای بهبود عملکرد از گسترش دایره واژگان برای واژه‌های نماینده استفاده کرده است. در این پژوهش همه سه‌حرفی‌های فارسی ساخته شده و تلفظ آن‌ها با روش‌های نویسه به واژ تولید شده است. این پژوهش بر روی دادگان فارس‌دات کوچک [43] انجام شده است و ۱۷۰ کلیدواژه تکواژه‌ای استفاده شده است که ۹۸/۷۲ کلیدواژه درون واژگان بازنیستی گفتار (INV)^۸ و ۴۹/٪ کلیدواژه OOV بوده‌اند. این سامانه به نرخ خطای واژه رسیده است. برای تشخیص عبارت‌ها با معیار ارزش وزنی واقعی عبارت (ATWV)^۹ عدد ۰/۰۲ با استفاده از واژه‌های نماینده و عدد ۰/۳ با گسترش واژگان بازنیستی گفتار حاصل شده است. با توجه به حجم کم دادگان فارس‌دات کوچک هم برای آموزش و هم برای آزمون و حجم بالای واژه‌های OOV (حدود ۴۲/٪) به نظر نتایج خوبی بهدست آمده است.

غدیری‌نیا در [37] بر اساس تشخیص واژی و واژه‌ای سامانه‌ای را برای تشخیص عبارت‌های گفتاری فارسی طراحی کرده است. در این سامانه اسناد گفتاری در دو سطح واژه و واژه‌گذاری می‌شوند و کلیدواژه بر روی هر دوی

¹ Bottleneck neural network² IN-Vocabulary (INV)³ Actual Term Weighted Value (ATWV)⁴ Maximum Term Weighted Value (MTWV)⁵ Levenshtein Distance⁶ Bidirectional Long Short-Term Memory (BLSTM)

تصمیم دوگانه (بله یا خیر) برای این که آیا سامانه رخداد فرضی را قبول دارد یا خیر؟ و یک امتیاز تشخیص که نشان‌دهنده میزان واقعی اتفاق عبارت احتمالی است. این تصمیم‌گیری، تصمیم‌گیری «واقعی» نامید می‌شود. در ادامه این بخش به روش پیشنهادی این مقاله برای تشخیص عبارت‌های گفتاری پرداخته می‌شود.



(شکل-۱): ساختار کلی سامانه‌های تشخیص عبارت‌های گفتاری بر پایه بازشناسی گفتار

(Figure-1): Overall Structure of an STD based on ASR

۲-۳- روشن پیشنهادی

در این مقاله برای تشخیص عبارت‌های گفتاری از سامانه بازشناسی گفتار استفاده شده است. در ابتدا سامانه بازشناسی گفتار برای زبان فارسی طراحی و پیاده‌سازی شده است. در ادامه مجموعه صوت‌ها و گفتارها توسط سامانه بازشناسی گفتار به متن بازشناسی می‌شوند.

در این مرحله از آنجایی که نخستین بهترین خروجی بازشناسی گفتار دارای خطا است، از مشبک جستجوی سامانه بازشناسی گفتار که گرافی از تمام واژه‌های احتمالی بازشناسی شده است، استفاده می‌شود؛ سپس برای نمایه کردن خروجی بازشناسی گفتار، مبدل‌های حالت محدود به کار رفته گرفته شده‌اند. این روش از سرعت و حجم محاسباتی قابل قبولی برخوردار است [46]. در مرحله جستجوی کلیدواژه، اگر کلیدواژه در دایره واژگان سامانه بازشناسی گفتار وجود داشته باشد، کلیدواژه به طور مستقیم بر روی نماینده جستجو می‌شود؛ اما اگر کلیدواژه از واژه‌های خارج از دایره واژگان سامانه بازشناسی گفتار باشد، از روش واژه‌های نماینده [31] استفاده می‌شود. در این روش از روی واژگان سامانه بازشناسی گفتار، واژه‌های مشابه آویی برای جستجوی کلیدواژه‌های OOV ساخته و از آن‌ها برای جستجو بر روی نماینده استفاده می‌شود. از آنجایی که در این

صرفجو در [39] با استفاده از سامانه بازشناسی گفتار بر روی اسناد گفتاری، بازیابی انجام داده است. در این پایان‌نامه با هدف مقابله با خطای خروجی سامانه بازشناسی گفتار، برای هر واژه مهم موجود در سند، تعدادی کلمه مشابه در نظر گرفته می‌شود. به این ترتیب بهازی هر سند، تعدادی سند جایگزین به دست می‌آید. با توجه به تعداد زیاد اسناد جایگزین، در مرحله بعد، مؤثرترین اسناد جایگزین، انتخاب و این اسناد جایگزین، به اسناد فعلی اضافه و درنهایت الگوریتم بازیابی روی مجموعه جدید اسناد اعمال می‌شود.

اخلاقی و سلطانی در [38]، [37] نیز بر پایه متن خروجی بازشناسی گفتار بر روی بازیابی اسناد گفتاری متوجه شده‌اند. در این پژوهش‌ها سعی در تشخیص خطای بازشناسی در واژه‌های بازشناسی شده و در نظر گرفتن واژه‌های مشابه برای واژه‌های دارای خطأ بوده است.

۳- روش پیشنهادی

۳-۱- مبانی تشخیص عبارت‌های گفتاری

سامانه‌های تشخیص عبارت‌های گفتاری، بر اساس بازشناسی گفتار از دو مرحله تشکیل شده‌اند؛ ۱- بازشناسی مجموعه سندهای صوتی و گفتاری و ۲- نمایه کردن و جستجو و نتایج بازشناسی گفتار. درواقع قبل از جستجو، باید مجموعه اسناد نمایه شوند و در مرحله نمایه کردن، اطلاعاتی از عبارت‌های مورد جستجو وجود ندارد و نمی‌توان از آن‌ها استفاده کرد. بخش‌های اصلی سامانه‌های تشخیص عبارت‌های گفتاری بر اساس سامانه‌های بازشناسی گفتار در شکل (۱) نشان داده شده است.

عبارت‌ها دنباله‌ای از واژه‌های گفتاری هستند که ارتباط زبانی با یکدیگر ندارند، بلکه گستره دستور زبانی دارند و می‌توانند از واژگان منفرد تا گروه واژه^۱ باشند. عبارت‌ها باید یک معنی به‌طور کامل قابل تشخیص داشته باشند بهصورتی که یک کاربر فرضی تمایل به پیدا کردن آن‌ها داشته باشد. عبارت‌ها شامل پنج یا کمتر واژه هستند. عبارت‌ها فقط بهصورت نوشتاری هستند و سامانه باید تفاوت تلفظی آن‌ها را در املای یکسان در نظر بگیرد.

برای هر رخداد محتمل از عبارت مورد جستجو، سامانه باید خروجی‌هایی شامل موارد زیر را گزارش کند: شروع و پایان زمان رخداد عبارت در سند مربوطه، یک

^۱ phrase



CD) پذیرش یا بازشناسی می‌کند.

^۱ Automata

روش کلیدواژه‌ها بازنمایی می‌شوند، دیگر نیازی به نمایه‌گذاری جدا نیست. با توجه به مشابهت آوایی و اجها روش‌های جستجوی کلیدواژه‌های OOV داری هستند که زیادی هستند، برای بهبود این روش از یک شبکه نادرست است [46]. البته به طور کلی این مسئله با جستجو بر روی آن جستجو انجام داد [47]. اینها به انتهای LSTM-CTC استفاده می‌شود [47]. این شبکه بر اساس بازشناسی نویسه به جای اج است و برای شناسایی OOV‌ها (به طور معمول آوای OOV‌ها موجود نیست ولی نوشتار آنها موجود است) مناسب‌تر است. همچنین این شبکه بر اساس واژه‌های جدا از هم (نه گفتار پیوسته) آموزش داده و باعث می‌شود اثر مدل زبانی را در بازشناسی کمتر می‌کند.

در ادامه به مبانی این روش‌ها پرداخته می‌شود. جزئیات این روش‌ها، طریقه پیاده‌سازی، الگوریتم‌ها و مشخصات محاسباتی آنها و نتایج بر روی دادگان مختلف در مقاله‌های مرجع آنها آمده است.

۳-۳- الگوریتم تشخیص

همان‌طور که اشاره شد، نخستین بهترین خروجی بازشناسی شده دارای خطای خطا است و این خطای خطا بر روی فرایند تشخیص عبارت‌ها تأثیرگذار است. همان‌طور که در [5] عنوان شده است، با سامانه بازشناسی گفتار با حدود بیست درصد خطای خطا می‌توان به دقت انسانی در بازیابی اطلاعات دست‌یافته. برای مقابله با عدم قطعیت در خروجی بازشناسی گفتار به جای جستجو بر روی نخستین بهترین خطای بازشناسی شده، بر روی مشبك خروجی آن جستجو انجام می‌شود. مشبك خروجی یک گراف‌های بدون دور جهت‌دار وزن‌دار است که همه احتمال‌های خروجی بازشناسی شده را نمایش می‌دهد. با توجه به حفظ تمام احتمالات در مشبك روش پیشنهادی باید از حجم و زمان محاسباتی معقولی برخوردار باشد.

برای تطبیق رشته در متن، از الگوریتم مشبك عامل [10] استفاده می‌شود. مراحل تولید این مشبك از روی یک ماشین خودکار^۱ مربوط به متن در [10] شرح داده شده است. این روش ماشین خودکار عامل نامیده شده است. ماشین خودکار عامل F(u) یک رشته مانند IR ، یک پذیرنده حالت محدود کمینه و قطعی است که مجموعه زیررشته‌های IR را عیناً و به درستی بازشناسی می‌کند. درواقع مشبك عامل یک رشته مانند ABCD هر زیررشته از آن را (مانند A, AB, CD)

در الگوریتم جستجو سعی بر این است که مشبك^۲ خروجی ASR که با یک ماشین خودکار نمایش داده می‌شود، به یک مبدل عامل تبدیل شود تا بتوان بر روی آن جستجو انجام داد [46]. البته به طور کلی این مسئله با جستجو بر روی متن متفاوت است؛ زیرا برخلاف متن، خروجی ASR به صورت غیرقطعی است. در این روش مبدل حالت محدود وزنی مشبك‌های همه گفته‌ها در مجموعه جستجو از مبدل‌های حالت محدود وزنی اختصاصی، به یک ساختار مبدل عامل تبدیل می‌شود؛ که در آن زمان شروع، زمان پایان و مشبك احتمالات پسین هر قطعه واژه را به صورت یک هزینه‌های سه‌بعدی ذخیره می‌کند. این مبدل عامل در واقع یک نمایه‌گذاری وارون از همه رشته‌واژگان در مشبك خروجی بازشناسی گفتار است.

بعد از نمایه‌کردن تمام استناد گفتاری، جستجوی درخواست کاربر بر روی آن انجام می‌شود. درخواست کاربر به طور معمول یک رشته‌ی بدون وزن است ولی می‌توان آن را به صورت یک ماشین خودکار وزنی دلخواه در نظر گرفته شود. این کار عبارت‌های منظم را که می‌توانند به ماشین خودکار تبدیل شوند؛ پوشش می‌دهد. با هم‌نهشت کردن^۳ ماشین خودکار وزنی درخواست با نمایه‌ها (که به صورت ماشین خودکار وزنی ذخیره شده‌اند) در قسمت ورودی جستجو انجام می‌پذیرد.

۴- واژه‌های نماینده

همان‌طور که بیان شد، یکی از چالش‌های تشخیص عبارت‌های گفتاری مبتنی بر بازشناسی گفتار کلیدواژه‌های خارج از واژگان (OOV) است که برای مقابله با این چالش، در [31] روش واژه‌های نماینده معرفی شده است. ایده اصلی این روش از روش گسترش پرسش در بازیابی متن گرفته شده است. در گسترش پرسش، یک پرسش را با واژه‌ها یا اصطلاحات مشابه آن پرسش در معنا یا شکل واژگانی مانند متراffها یا انواع ساخت واژی^۴ تکمیل و تقویت و بر اساس همین ایده، برای جستجوی یک پرسش، آن را با کلیدواژه‌های مشابه آوایی آن جستجو می‌کنند. به این واژه‌های مشابه آوایی، واژه‌های نماینده می‌گویند. روش واژه‌های نماینده، روشی بر اساس مبدل‌های حالت محدود وزنی هستند که در آن مبدل‌های مشابه آوایی برای کلیدواژه OOV از روی INV ساخته می‌شود.

² Lattice

³ compos

⁴ Morphological

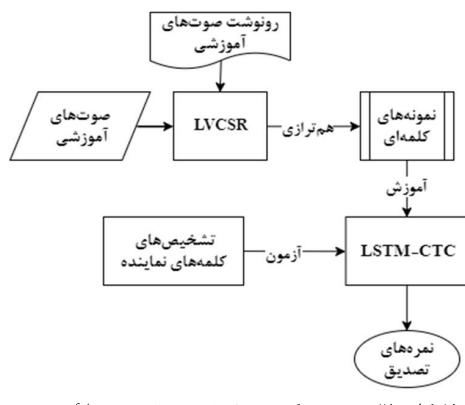
جامع آوایی برای تصدیق خروجی روش واژه‌های نماینده است. روش تصدیق، ویژگی قطعه‌ها را متناسب با دنباله حروف همتراز می‌کند و احتمال پیشین آن را به عنوان امتیاز اطمینان در نظر می‌گیرد.

در [47] نشان داده شده است که شبکه‌های دسته‌بندی زمانی پیوندگرا (CTC) بسیار مناسب برای برچسبزنی این‌چنین دنباله داده‌های بدون قطعه‌بندی هستند. با درنظر گرفتن یک دنباله ویژگی ورودی X به طول T و دنباله حروف آن W , CTC یکتابع هدف است که به‌شکل زیر تعریف می‌شود:

$$L_{CTC}(X, W) = \sum_{C_W} p(C|X) = \sum_{C_W} \prod_{t=1}^T p(c_t|X) \quad (2)$$

که C_W هر دنباله برچسب با طول T متناسب با دنباله درست حروف W است. با جمع‌بستن بر روی تمامی مجموعه محل‌های برچسب که دنباله یکسان W را حاصل می‌کند، CTC یک توزیع احتمالی بر روی برچسبزنی‌های ممکن مشروط به دنباله ورودی X , را تعیین می‌کند.

در این روش، یک لایه CTC در بالای شبکه عصبی LSTM قرار داده شده است که به طور مستقیم با استفاده از ویژگی گفتار اصلی آموزش داده می‌شود. در مرحله آموزش، نمونه‌های واژه‌ای (نه به صورت جمله پیوسته) به یک شبکه چندلایه LSTM که از معیار CTC استفاده می‌کند، داده می‌شود. برای آماده کردن نمونه‌های واژه‌ای از صوت‌ها، از سامانه بازشناسی گفتار برای همترازی واژه‌ها در سطح گفته‌ها استفاده می‌شود. در مرحله آزمون، نمونه‌های کشف شده واژه‌های نماینده به یک شبکه عصبی LSTM وارد می‌شود و امتیاز تصدیق خروجی هستند. سامانه تصدیق LSTM-CTC در شکل (۲) نشان داده شده است.



(شکل-۲): تصدیق کننده LSTM-CTC. مرحله آموزش و آزمون [47].

(Figure-2): LSTM-CTC verification. Train and test phases [47].

فرض کنید K بیانگر یک پذیرنده حالت محدود برای یک کلیدواژه OOV و L_2 یک مبدل حالت محدود برای تلفظ کلیدواژه OOV باشد؛ E را یک مبدل ویرایش فاصله فرض کنید که یک رشته واژ را به هر دنباله واژی دیگر با هزینه‌های تخمینی از یک ماتریس درهم‌ریختگی نگاشت می‌کند؛ همچنین L_1 را واژگان سامانه بازشناسی گفتار در نظر بگیرید؛ روش WFST برای تولید کلیدواژه نماینده K با عبارت زیر توصیف می‌شود:

$$K' = \text{Project}(\text{ShortestPath}(K \circ L_2 \circ E \circ (L_1^*)^{-1})). \quad (1)$$

در روش جستجوی کلیدواژه‌های dNV عبارت جستجو به صورت یک مبدل حالت محدود بازنمایی می‌شد و با همراهشستی بر روی نمایه‌ها (ذخیره شده به صورت یک مبدل حالت محدود) جستجو انجام می‌گرفت. در روش واژه‌های نماینده، عبارت شماره (۱) مبدل حالت محدود وزنی واژه OOV را می‌سازد. این مبدل حالت محدود وزنی یک فاصله‌یاب بین واژه‌های داخل واژگان ASR و آواز تولیدی واژه OOV است، برای تخمین فاصله از ماتریس درهم‌ریختگی واژه‌ها استفاده می‌شود.

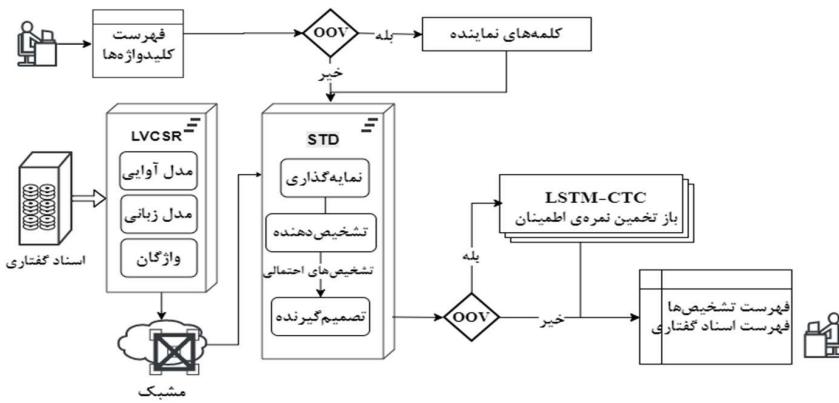
یکی از مزیت‌های این روش، عدم ذخیره نمایه‌گذاری جدالگانه (مثل نمایه‌های واژی) برای کلیدواژه‌های OOV است، این کلیدواژه‌ها از نمایه‌های کلیدواژه‌های IV استفاده می‌کنند.

یکی از ضعف‌های عمومی روش‌های مرتبط با OOV‌ها بالای بودن هشدار نادرست در خروجی آن‌ها است. هشدارهای نادرست به علت متكلی بودن این روش‌ها به بازشناسی واژ است. برای همین روش‌هایی برای مقابله با این مشکل ارائه شده است [48] که در زیر یک روش برای بازتخمین خروجی واژه‌های نماینده بیان می‌شود.

۱-۴-۳- بهبود عملکرد روش واژه‌های نماینده

روش‌های جستجو برای عبارت‌های OOV به طور معمول دارای هشدار نادرست زیادی هستند. در اینجا یک روش برای بهبود عملکرد واژه‌های نماینده ارائه شده است. ایده اصلی این روش بازتخمین امتیاز خروجی‌های روش واژه‌های نماینده با استفاده از یک شبکه حافظه کوتاه‌مدت ماندگار و دسته‌بندی زمانی پیوندگرا (LSTM-CTC) است [47].

هدف اصلی در این روش، محاسبه دوباره امتیاز اطمینان برای خروجی واژه‌های نماینده است. ایده اصلی این روش استفاده از یک شبکه LSTM-CTC به عنوان یک مدل



(شکل-۳): روند نمای کلی سامانه تشخیص عبارت‌های گفتاری پیشنهادی

(Figure-3): Proposed flowchart for spoken term detection

واژگان سامانه بازشناسی گفتار انتخاب می‌شود. واژه‌های انتخاب شده به جای کلیدواژه خارج از دادگان بر روی نماینده جستجو می‌شود.

یکی از مزیت‌های روش واژه‌های نماینده عدم نیاز به نمایه دیگری مانند نمایه واجی است. در این روش واژه‌های نماینده به جای کلیدواژه‌های خارج از واژگان سامانه بازشناسی، مستقیم بر روی نمایه‌های واژه جستجو می‌شوند. این مزیت باعث کم شدن حجم محاسبات و عدم نیاز به ذخیره نمایه جداگانه می‌شود.

از آنجایی که روش واژه‌های نماینده از شباهت واجی بین کلیدواژه خارج از واژگان و واژگان سامانه بازشناسی استفاده می‌کند، به علت شباهت آوازی بین برخی واجه، دارای هشدارهای نادرست زیاد است. برای مقابله با این هشدارهای نادرست یک شبکه LSTM-CTC برای تخمین دوباره امتیاز اطمینان خروجی واژه‌های نماینده پیشنهاد شده است. یکی از مزیت‌های شبکه LSTM-CTC بازشناسی گفتار بر اساس نویسه به جای واج است. از آنجایی که تلفظ واژه‌های خارج از واژگان در دسترس نیست و با استفاده از شبکه‌های نویسه به واج ایجاد می‌شود، به طور معمول این تلفظها دارای خطأ هستند؛ بنابراین بازشناسی نویسه در شبکه‌های LSTM-CTC این خطأ را از بین می‌برد.

علاوه بر مشکل شباهت واجی در روش واژه‌های نماینده، عدم امتیازدهی به واژه‌های خارج از دادگان در رمزگشایی بازشناسی گفتار، مشکل زا است. در رمزگشایی سامانه بازشناسی گفتار هر دو، وزن مدل آوازی و وزن مدل زبانی مهم هستند. به طور معمول وزن واژه‌های خارج از دادگان در مدل زبانی حذف و این باعث می‌شود، برخلاف داشتن وزن مدل آوازی برای واژه‌های خارج از دادگان، ولی در مدل زبانی وزنی نداشته باشدند.

۳-۵- شماتی کلی روش

شماتی کلی روش پیشنهادی در شکل (۳) آمده است. این سامانه از دو زیرسامانه بازشناسی گفتار و تشخیص عبارت‌ها تشکیل شده است. ابتدا سامانه بازشناسی گفتار اسناد گفتاری ورودی را به مجموعه‌ای واژگان که در قالب شبکه‌ها نگهداری می‌شوند، تبدیل می‌کند. شبکه‌ای حاصل، به عنوان ورودی به زیرسامانه تشخیص وارد و در این زیرسامانه ابتدا شبکه‌ها نمایه‌گذاری می‌شوند. برای این کار، تمام واژه‌هایی که در این شبکه‌ها بازشناسی شده‌اند، به صورت یک سه‌تایی (احتمال پسین، زمان شروع، زمان پایان) نمایه می‌شوند. احتمال پسین به عنوان امتیاز اطمینان نمایه‌ها استفاده می‌شود.

در بخش جستجو از دو روش متفاوت برای جستجوی کلیدواژه‌های درون واژگان و خارج از واژگان استفاده می‌شود. اگر کلیدواژه ورودی در واژگان سامانه بازشناسی گفتار وجود داشته باشد، این کلیدواژه مستقیم بر روی نمایه‌ها جستجو و تشخیص‌های احتمالی به تصمیم‌گیرنده وارد می‌شود. تصمیم‌گیرنده بر اساس امتیازهای نمایه‌ها و یک مرز تصمیم، تشخیص‌های احتمالی را پذیرش یا رد می‌کند.

در صورتی که کلیدواژه در واژگان سامانه بازشناسی وجود نداشته باشد، با استفاده از واژه‌های نماینده مدل‌های حالت محدود وزنی، مشابه آوازی آن از روی واژگان بازشناسی گفتار ساخته می‌شود و این مدل‌های نماینده بر روی نمایه‌ها مورد جستجو قرار می‌گیرد. در این روش ابتدا تلفظ کلیدواژه خارج از واژگان ایجاد، سپس با استفاده از ماتریس درهم‌ریختگی واج، شباهت آوازی بین تلفظ کلیدواژه و واژگان سامانه بازشناسی محاسبه و شبیه‌ترین واژه‌ها به



پایگاه داده فارس‌دات بزرگ: فارس‌دات بزرگ در مجموع ۱۴۰ ساعت سیگنال گفتار است که تقطیع و برچسب‌گذاری در حد واژه شده‌اند. تعداد ۱۰۰ گوینده ۳۹ زن و ۶۱ مرد) با لهجه رایج فارسی و هر گوینده ۲۵-۲۰ صفحه متن با موضوعات مختلف را خوانده‌اند. برخلاف فارس‌دات کوچک که در محیط ساکت و بدون اعوجاج^۲ ضبط شده است، فارس‌دات بزرگ در محیط اداری و دفتر کار ضبط شده است. چهار میکروفون، یک میکروفون رومیزی یکسویه، دو میکروفون یقمه‌ای و یک میکروفون گوشی برای ضبط سیگنال گفتار استفاده شده است. همه سیگنال‌های گفتاری همزمان با دو میکروفون ضبط شده‌اند؛ میکروفون رومیزی در ضبط همه سیگنال‌ها استفاده شده و سه میکروفون دیگر هر کدام در یک‌سوم دادگان استفاده شده است.

دادگان اخبار فارسی صداوسیمای جمهوری

اسلامی ایران: این پایگاه داده حدود سی ساعت اخبار فارسی پخش شده از سیمای جمهوری اسلامی ایران را در حد جمله تقطیع و برچسب‌گذاری کرده است. این دادگان شبیه واقعیت و آن چیزی است که در سازمان صداوسیما تولید می‌شود. در این دادگان اخبار فارسی در جملات کمتر از یک دقیقه بخش‌بندی و فایل‌های صوتی آن جدایگانه به‌همراه رونوشت متن آن تهیه شده است.

۴-۲- سامانه بازشناسی گفتار فارسی

سامانه‌های بازشناسی گفتار با دایره لغات وسیع، به طور معمول از سه مدل مختلف تشکیل شده‌اند: مدل آوایی، مدل زبانی و واژگان. در زیر به‌طور جداگانه حجم دادگان برای هر سه مدل توضیح داده می‌شود.

۴-۲-۱- مدل آوایی

پایگاه داده فارس‌دات کوچک: در آزمایش‌های مرتبط با این دادگان، از ۲۲۴ گوینده برای آموزش مدل آوایی، ۵۰ گوینده برای توسعه و ۳۰ گوینده برای آزمون استفاده شده است [۵۰].

پایگاه داده فارس‌دات بزرگ: در این پژوهش از هفتاد ساعت دادگان میکروفون رومیزی فارس‌دات بزرگ استفاده شده است. در این پژوهش، هفتاد درصد این دادگان برای آموزش مدل آوایی، هفتاد درصد برای توسعه و

در این روش، شبکه LSTM-CTC به صورت واژه به واژه (و نه به صورت پیوسته) آموزش داده می‌شود. به علت آموزش واژه به واژه، وزن مدل زبانی (دبالة واژه‌ها) در مرحله رمزگشایی تأثیر بهسزایی ندارد و بیشتر مدل آوایی در آن تأثیرگذار است. تأثیر کم‌وزن مدل زبانی در رمزگشایی، یکی دیگر از مزیت‌های این روش آموزش است. در اینجا بیشتر امتیاز مدل آوایی تأثیرگذار است.

در مرحله آزمون، تشخیص‌های روش واژه‌های نماینده برای بازشناسی به شبکه LSTM-CTC داده می‌شود و شبکه این تشخیص‌ها را دوباره بازشناسی می‌کند و برای آن‌ها امتیاز اطمینان دوباره محاسبه می‌شود. این امتیاز اطمینان به عنوان امتیاز جدید تشخیص‌ها استفاده و به تصمیم‌گیرنده داده می‌شود. تصمیم‌گیرنده بر اساس مرز تصمیم درباره پذیرش یا رد آن‌ها تصمیم‌گیری می‌کند.

۴- ارزیابی و نتایج

۴-۱- دادگان

در این پژوهش تمرکز بر روی زبان فارسی است و از سه پایگاه داده استفاده شد. پایگاه داده فارس‌دات کوچک [۴۳]، فارس‌دات بزرگ [۴۴] و پایگاه داده دادگان اخبار فارسی صداوسیمای جمهوری اسلامی ایران^۱ در این پژوهش است.

پایگاه داده فارس‌دات کوچک: در این دادگان، ۳۰۴ گوینده به صورت اتفاقی بیست جمله را می‌خوانند. در مجموع ۴۵۰ جمله استفاده شده است، که هر گوینده هجده جمله را به صورت اتفاقی می‌خواند و دو جمله را همه گویندگان خوانده‌اند. جمله‌ها مصنوعی ساخته شده‌اند به‌دلیل این‌که تنوع آوایی زبان فارسی پوشش داده شود. گویندگان از مناطق مختلف ایران انتخاب شده‌اند و در مجموع، دادگان از ده لهجه عمومی فارسی تشکیل شده است. نسبت مردان به زنان ۲:۱ است. دادگان در محیط کمونوفه با میانگین سیگنال به نویه حدود ۳۱ dB و نرخ نمونه‌برداری ۲۲۰۵۰ HZ ضبط شده‌اند. این دادگان در سطح واج تقطیع و برچسب‌گذاری شده‌اند. فارس‌دات کوچک را می‌توان با دادگان TIMIT [۴۹] در زبان انگلیسی مقایسه کرد. این دادگان در مجموع حدود پنج ساعت است.

^۱ این دادگان از طریق دکتر باقر باعلی عضو هیئت علمی دانشگاه تهران، دانشکده دانشکده ریاضی، آمار و علوم رایانه تهیه شد.

مقاله مرجع خاصی برای ارجاع دهی ندارد.

² Reverberation



بیست درصد برای آزمون انتخاب شدند. آموزش شامل سی گوینده زن و ۴۳ گوینده مرد و حدود ۵۳ ساعت سیگنال گفتار؛ توسعه شامل پنج گوینده زن و یازده گوینده مرد و حدود یازده ساعت سیگنال گفتار و آزمون شامل چهار گوینده زن و هفت گوینده مرد و حدود هشت ساعت سیگنال گفتار است.

دادگان اخبار فارسی صداوسیمای جمهوری اسلامی ایران: این دادگان تجاری است و بخشی از آن با همکاری آقای دکتر باباعلی عضو هیئت‌علمی دانشگاه تهران فقط به عنوان آزمون در حدود سه ساعت در اختیار این پژوهش قرار گرفت. این دادگان شامل تمامی بخش‌های اخبار شامل گفتار رسمی گوینده، گفتارهای محاوره‌ای گزارش‌ها، محیط‌ها و لهجه‌های مختلف و گفتارها با و یا بدون موسیقی است. این دادگان شامل ۸۶۱ گفته گفتاری مجزا است.

در جدول (۲) خلاصه اطلاعات آوایی دادگان و تقسیم‌بندی آن‌ها برای آموزش مدل آوایی آمده است.

۴-۲-۲- واژگان و آموزش مدل زبانی

پیکره آموزشی مدل زبانی و واژگان مرتبط برای هر دادگان به شرح زیر است:

فارس‌دات کوچک: از تمامی ۴۵۰ جمله منحصر به‌فردی که در این دادگان به‌کاررفته رفته، برای ساخت مدل زبانی موردنیاز بازشناسی گفتار استفاده شده است. این دادگان شامل ۱۴۱۴ واژه یکتا هستند که معادل آوایی آن‌ها در دادگان موجود است.

فارس‌دات بزرگ: این دادگان شامل حدود ۷۵۰ هزار واژه و دارای ۳۸۷۵ واژه یکتا است. متأسفانه برخلاف فارس‌دات کوچک، فارس‌دات بزرگ دارای واژگان و لغتنامه آوایی از لغات نیست. درواقع برچسب‌های واژه‌ای در این دادگان به صورت آوایی نوشته شده‌اند و معادل فارسی آن‌ها موجود نیست. با وجود صرف وقت، تهیه لغتنامه از روی برچسب‌های آوایی و معادل‌سازی فارسی آن‌ها کار زمان‌بری است؛ بنابراین در اینجا از همین برچسب‌های آوایی به عنوان واژگان بازشناسی گفتار استفاده شد. از جملاتی که مشکل از همین آواها بودند، مدل زبانی بازشناسی گفتار ساخته شد. پیکره آموزشی این مدل زبانی شامل حدود ۱۷۰۰ جمله و حدود پانصد هزار واژه که شامل متن‌های آموزش و توسعه است.

اخبار فارسی صداوسیمای جمهوری اسلامی ایران: در این حالت، از دو پیکره متنی برای ساخت مدل زبانی استفاده شد: پیکره همین دادگان و پیکره ایستنا. پیکره متنی ایستنا^۱ شامل حدود ۵۵۰ میلیون واژه و شامل اخبار خبرگزاری ایستنا است. پیکره دادگان اخبار فارسی از متن همین اخبار و شامل حدود ۳۰ هزار واژه است. علت انتخاب دو پیکره متنی برای ساخت مدل زبانی، تأثیر آن در رمزگشایی بازشناسی گفتار است؛ از آنجایی که اگر از متن مرتبط با دادگان گفتاری به عنوان پیکره متنی استفاده شود، مدل بیش‌برازش^۲ می‌شود؛ از پیکره متنی جدا استفاده شد.

برای تهیه لغتنامه آوایی این دادگان، حدود شش هزار واژه یکتا به‌کاررفته در آن را استفاده کردیم. از آنجایی که لغتنامه آوایی واژه‌های این دادگان وجود نداشت، ساخت این لغتنامه به صورت دستی انجام و برای ساخت واژگان این دادگان، از واژگان زیایی فارسی [۵۱] استفاده شد. واژگان زیایی زبان فارسی، شامل حدود ۵۵ هزار مدخل که هر مدخل دارای اطلاعات مربوط به صورت نوشتاری واژه در خط فارسی، ساخت واژی، مقوله واژگانی، الگوی تکیه و بسامد واژه است. برای تهیه واژگان زیایی، یک پیکره متنی ۵۵ میلیون واژه‌ای ملاک استخراج واژه‌ها قرار گرفته است. واژگان زیایی از حدود صدهزار واژه با بسامدهای متفاوت تشکیل شده است. بعد از حذف صورت‌های تصریفی از فهرست بالا، حدود ۴۴ هزار واژه به مفهوم علمی آن به دست آمد؛ همچنین برای غنی‌سازی فهرست مداخل به‌ نحوی که شامل واژه‌های عامیانه و برخی واژه‌های علمی باشد، از فرهنگ فارسی امروز استفاده شده است. روند کار به این صورت بود که برخی از واژگان یکتای استفاده شده در دادگان اخبار فارسی در واژگان زیایی وجود داشت؛ دیگر واژه‌ها با استفاده از کتابخانه «هضم» [۵۲] ریشه‌یابی^۳ شدند و اگر تلفظ ریشه‌ها و وندها در واژگان زیایی بودند به هم پیوند می‌خوردند تا تلفظ واژه پیدا شود. تلفظ برخی از واژه‌ها با هیچ کدام از این راه‌کارها ساخته نشد؛ بنابراین به صورت دستی تلفظ آن‌ها نوشته شد.

در جدول (۳) خلاصه‌ای از اطلاعات متنی و واژگان دادگان آمده است.

برای آموزش مدل زبانی ابزار IRSTLM [۵۳] به کار گرفته و از مدل زبانی سه‌تایی استفاده شد.

¹ ISNA

² Overfit

³ Stemmer



softmax است و درنهایت نتایج گزارش و با هم مقایسه شده‌اند.

۴-۴-۴- معیارهای ارزیابی

۱-۴-۴-۱- ارزیابی سامانه بازشناسی گفتار

در بازشناسی گفتار سه خطای حذف^۱، درج^۲ و جایگزینی^۳ اتفاق می‌افتد. خطای جایگزینی وقتی است که یک واژه به جای یک واژه دیگر بازشناسی شود. خطای حذف، بیان‌گر حالتی است که یک واژه که در گفتار وجود دارد، در خروجی بازشناسی حذف شود و خطای درج زمانی رخ می‌دهد که یک واژه که در گفتار وجود ندارد، در خروجی بازشناسی اضافه شود. برای ارزیابی بازشناسی گفتار از معیار ترتیخ طایی واژه (WER)^۴ استفاده شد که صورت این کسر بیان‌گر جمع این سه نوع خطا است:

$$WER = \frac{(N^{Del} + N^{ins} + N^{sub})}{N^{all}} \quad (۳)$$

۱-۴-۴-۲- ارزیابی سامانه تشخیص عبارت‌های گفتاری

دو معیار ارزیابی سامانه‌های تشخیص عبارت‌های گفتاری در سال ۲۰۰۶ [۳] معرفی شدند و هم‌اکنون نیز در معیارهای NIST استفاده می‌شوند [۲۹]. معیار گرافیکی به‌وسیله منحنی‌های مصالحه تشخیص (DET) ^۵ [۵۴] و برای یک نقطه عملیاتی خاص در فضای منحنی DET از ارزش وزنی عبارت (TWV)^۶ استفاده می‌شود. اولی یک نمایش شهودی از عملکرد سامانه برای کاربردهای با دقت بالا یا فراموشی بالا فراهم می‌کند. درحالی‌که TWV برای توسعه‌دهنگان یک معیار عملکرد واحد به عنوان هدفی برای بهینه‌سازی سامانه ایجاد می‌کند.

در تشخیص عبارت‌های گفتاری، یک رخداد فرضی به عنوان تشخیص نامیده می‌شود؛ اگر تشخیص مربوط به رخداد درست باشد، یک موفقیت^۷ نامیده می‌شود، در غیر این صورت یک هشدار نادرست^۸ است. اگر رخداد درست تشخیص داده نشود، از دست‌رفته^۹ نامیده می‌شود. با توجه به

¹ Deletion

² Insertion

³ Substitution

⁴ Word Error Rate (WER)

⁵ Detection Error Tradeoff (DET) curves

⁶ Term-Weighted Value (TWV)

⁷ Hit

⁸ False alarm

⁹ Miss

(جدول-۲): خلاصه اطلاعات دادگان. #spk تعداد گویندگان، T

مدت زمان به ساعت.

(Table-2): summary of databases. #spk Number of speakers, T length in hours.

دادگان	کل						آزمون	
	T	#spk	T	#spk	T	#spk		
فارس‌داد کوچک	0. 2	30	0. 4	50	3	224	3. 5	30 4
فارس‌داد بزرگ	8	11	11	16	5 3	73	71	10 0
اخبار فارسی صداوسیما	3	-	-	-	-	-	3	-

(جدول-۳): خلاصه اطلاعات پیکره متنی دادگان.

(Table-3): summary of text corpus of datasets

دادگان	دادگان	تعداد واژه‌های یکتا	تعداد واژه‌های
فارس‌داد کوچک	1.4K	3 K	
فارس‌داد بزرگ	3.9 K	750 K	
اخبار فارسی صداوسیما	6 K	30 K	
ایستنا	990 K	550 M	

۴-۳- سامانه تشخیص عبارت‌های گفتاری

همان‌طور که در قبل گفته شد، یکی از چالش‌های استفاده از بازشناسی گفتار برای تشخیص عبارت، تشخیص کلیدواژه‌های خارج از واژگان است. این واژگان به‌هیچ‌وجه در خروجی بازشناسی نمایش داده نمی‌شوند و با روش‌های مبتنی بر تشخیص کلیدواژه‌های درون واژگان نمی‌توان آن‌ها را تشخیص داد. برای همین باید الگوریتم‌های متفاوتی برای تشخیص واژگان داخل واژگان و خارج از واژگان استفاده کرد. درواقع سامانه ارائه شده در این پژوهش، از دو الگوریتم جدا برای تشخیص این دو نوع کلیدواژه استفاده می‌کند.

برای کلیدواژه‌های خارج از واژگان از روش واژه‌های نماینده استفاده شد. یکی از مزیت‌های این روش، نیاز نبودن به نمایه کردن دوباره استناد گفتار است و از نمایه مربوط به واژه‌های درون واژگان استفاده می‌کند که برای بهبود روش واژه‌های نماینده، از یک شبکه LSTM-CTC استفاده شد. آموزش این شبکه با ورودی‌های واژه به واژه به جای گفتار پیوسته است. برای آموزش، از دادگان فارس‌داد بزرگ با پانصدهزار واژه استفاده شده است. برای ارزیابی بهتر، شبکه LSTM مورد استفاده دارای یک تا پنج لایه دوسویه و ۱۵۰ سلول حافظه در هر لایه LSTM و دارای یک لایه پایانی

است. سامانه‌ای که هیچ خروجی نداشته باشد TWV برابر صفر دارد و مقادیر منفی برای TWV هم ممکن هست. ارزش وزنی واقعی عبارت (ATWV)^۲: در حالی که منحنی‌های DET عملکرد را برای تمامی مقادیر ممکن θ نشان می‌دهند، دو نقطه از منحنی DET موردنوجه هستند؛ زیرا تعیین می‌کنند که آستانه تصمیم واقعی سامانه بهینه است یا نه. اولی، ارزش وزنی واقعی عبارت (ATWV) است که TWV با استفاده از تصمیم‌های واقعی محاسبه می‌شود. دومی، ارزش بیشینه وزنی عبارت (MTWV)^۳ است. MTWV مقدار TWV در نقطه‌ای از منحنی DET است که یک مقدار از θ که TWV را بیشینه می‌کند. اختلاف بین مقادیر ATWV و MTWV نشان‌دهنده مزایای انتخاب آستانه تصمیم‌گیری واقعی بهتر است. معیار اصلی عملکرد سامانه تشخیص عبارات گفتاری، ATWV است که در اینجا از آن استفاده شده است.

۴-۵- نتایج سامانه بازشناسی گفتار

در این پژوهش از سامانه متن باز کلیدی [48] به عنوان بستر بازشناسی گفتار استفاده شده است. دو سامانه بازشناسی گفتار به صورت جداگانه، یکی برای دادگان فارس‌داد کوچک و یکی برای دادگان فارس‌داد بزرگ آموزش داده و ارزیابی شد. برای دادگان اخبار فارسی از مدل آوای فارس‌داد کوچک و بزرگ به صورت جداگانه استفاده شد. در رمزگشایی بازشناسی گفتار وزن مدل زبانی نقش دارد؛ ازان‌جایی که متن‌های آموزشی مدل زبانی محدود بودند و این بر روی نتایج مؤثر است، بهترین نتیجه‌ها با وزن مشخص شده مدل زبانی در بازه ۱ تا ۱۷ در جدول (۴) آمده است. برای ارزیابی بهتر مدل آوای، نتایج با وزن یک، مدل زبانی در جدول (۵) آورده شده است. در این جدول‌ها مدل‌های آوای HMM تکواج (mono)، سه‌واجی با ویژگی‌های دلتا و دلتا-دلتا (tri1)، سه‌واجی به همراه MLLT^۵(tri2)، LDA⁴ + MLLT⁵(tri3)، سه‌واجی به همراه (tri3)، LDA + MLLT + SAT⁶(tri3) و SGMM⁷ گزارش شده‌اند.

² Actual Term Weighted Value (ATWV)

³ Maximum Term Weighted Value (MTWV)

⁴ Linear Discriminant Analysis

⁵ Maximum Likelihood Linear Transform

⁶ Speaker Adaptive Training

⁷ Subspace Gaussian Mixture Models

اینکه این سامانه‌ها، سامانه‌های تشخیصی هستند، به حد آستانه تشخیص وابسته هستند. ارزش وزنی عبارت (TWV) برابر با یک منهای ارزش میانگین از دست داده شده سامانه در هر عبارت است؛ بنابراین به صورت زیر تعریف می‌شود:

$$TWV(\theta) = 1 - \frac{\text{average}_{term}\{P_{Miss}(term, \theta) + \beta \cdot P_{FA}(term, \theta)\}}{(4)$$

$$\beta = \frac{C}{V} (Pr_{term}^{-1} - 1) \quad (5)$$

که در آن

$$P_{Miss}(term, \theta) = 1 - \frac{N_{correct}(term, \theta)}{N_{true}(term)} \quad (6)$$

$$P_{FA}(term, \theta) = \frac{N_{spurious}(term, \theta)}{N_{NT}(term)} \quad (7)$$

و $N_{correct}(term, \theta)$ تعداد درست (واقعی) تشخیص‌های term با یک امتیاز تشخیص مساوی یا بزرگ‌تر از θ ؛ $N_{spurious}(term, \theta)$ تعداد جعلی (غیرواقعی) تشخیص‌های term با یک امتیاز تشخیص مساوی یا بزرگ‌تر از θ ؛ $N_{true}(term)$ تعداد واقعی وقوع term در دادگان؛ $N_{NT}(term)$ تعداد فرست‌ها برای تشخیص غیرواقعی term در دادگان.

از آنجاکه هیچ ویژگی مجزا از «آزمایش‌ها» وجود ندارد؛ تعداد آزمایش‌های بدون هدف برای یک term $N_{NT}(term)$ تا حدی قراردادی تعریف می‌شود تا با تعداد ثانیه‌های گفتار در دادگان آزمون متناسب باشد؛ به خصوص:

$$N_{NT}(term) = n_{tps} \cdot T_{speech} - N_{true}(term) \quad (8)$$

که n_{tps} تعداد آزمایش‌ها در ثانیه‌ی گفتار است (به صورت دلخواه برابر با یک تنظیم می‌شود). وزن β هم برای احتمال پیشین یک عبارت و هم برای وزن‌های نسبی برای هر نوع خطأ در نظر گرفته می‌شود.

برای این ارزشیابی، نسبت هزینه ارزش c/v برابر ۰/۱ است؛ بنابراین ارزش از دست رفته در برابر یک هشدار نادرست ده برابر ارزش و احتمال پیشین یک term برابر $4^{-\alpha}$ است.

بیشینه ممکن برای TWV برابر یک (۱/۰) متناظر با سامانه «تمام‌عیار» یعنی بدون هشدار نادرست و بدون miss

¹ Cost/value ratio



84/21	mono		
92/93	tri1	فارس دات	
99/47	tri2	کوچک	
60/84	tri3		
49/8	SGMM		خبر فارسی
68/71	mono		
81/44	tri1	فارس دات	
76/39	tri2		
60/70	tri3	بزرگ	
35/37	SGMM		

به طور کلی آموزش مدل آوازی مخصوص یک گوینده نتایج بسیار بهتری نسبت به مدل آوازی مستقل از گوینده دارد. آموزش مدل آوازی نیاز به تعداد بالای ساعت گفتاری دارد و جمع آوری این تعداد ساعت برای همه گویندها مقدور نیست؛ همچنین در برخی کاربردها از یک گوینده فقط مقدار کمی داده موجود است؛ بنابراین به مدل‌های آوازی مستقل از گوینده که بتوان با مقدار کمی داده با گوینده منطبق بشوند، نیاز است. ازانجایی که داده‌های آموزشی از گویندهای مختلف در دادگان کم هستند، آموزش مدل‌های آوازی به طور معمول مستقل از گوینده آموزش داده می‌شوند و در مرحله آزمون روش‌های تطبیق گوینده مانند MLLT و SAT استفاده می‌شود [55]. ایده اصلی در MLLT استفاده از یک مبدل خطی هست؛ این مبدل خطی با بسیاری از گوسی‌ها مشترک هست و برای تغییر پارامترهای گوسی مدل آوازی از آن استفاده می‌شود. ازانجایی که این مبدل‌های خطی مشترک هستند، مقدار کمی از پارامترها نیاز به تخمین هستند.

در روش آموزشی بر پایه HMM در حالت‌های GMM، HMM، HMM‌ها به صورت مستقل در نظر گرفته می‌شود؛ در روش SGMM گوسی مشترک بین حالت‌ها در نظر گرفته می‌شود [56]. در روش SGMM یک مدل مخلوط جهانی بر مبنای کل گفتار آموزش داده می‌شود؛ سپس با تطبیق آن به مدل‌های مربوط به هر حالت مرسیم. آموزش با حجم داده‌های کم یکی از مزیت‌های این روش است.

با توجه به محدودبودن حجم دادگان در دسترس، در این پژوهش بر روی روش‌های یادگیری عمیق و شبکه‌های عصبی تمرکز نشد و از HMM استفاده شده است. با توجه به نتایج، علت خطاًی کم در دادگان فارس دات بزرگ، استفاده از مدل زبانی فقط بر مبنای دادگان آنها و همچنین شباهت و همگنی دادگان آزمون و آموزش این دو دادگان است. علت خطاًی بالا روی دادگان اخبار فارسی، درهمبودن آنها و تفاوت با دادگان آموزش

(جدول-۴): بهترین نتایج خطای بازناسی گفتار بر مبنای معیار WER با مدل زبانی منطبق با دادگان و وزن آن. HMM تکواج (mono)، سهواجی با ویژگی‌های دلتا و دلتا-دلتا (tri1)، سهواجی به همراه LDA + MLLT(tri2) و SGMM و LDA + MLLT + SAT(tri3)

(Table-4): Best results of speech recognition based on WER using language model matched to text of databases. Monophone(mono), triphone with delta and delta-delta (tri1), triphone with LDA+MLLT(tri2), triphone with LDA + MLLT + SAT(tri3), SGMM.

دادگان آزمون	مدل آوازی	وزن مدل زبانی	بهترین WER
فارس دات	mono	16	7/93
	tri1	17	9/72
	tri2	17	11/83
	tri3	17	11/35
	SGMM	17	9/55
فارس دات	mono	11	9/46
	tri1	17	4/41
	tri2	17	4/7
	tri3	17	3/42
	SGMM	15	2/71
کوچک	mono	13	77/12
	tri1	17	77/23
	tri2	17	76/48
	tri3	17	51/52
	SGMM	17	34/40
فارس دات	mono	14	54/79
	tri1	9	79/69
	tri2	11	71/77
	tri3	10	59/29
	SGMM	13	28/23
فارسی	mono	14	77/12
	tri1	9	79/69
	tri2	11	71/77
	tri3	10	59/29
	SGMM	13	28/23

(جدول-۵): نتایج خطای بازناسی گفتار بر مبنای معیار WER با مدل زبانی منطبق بر دادگان و وزن یک. HMM تکواج (mono)، سهواجی با ویژگی‌های دلتا و دلتا-دلتا (tri1)، سهواجی به همراه LDA + MLLT(tri2) و LDA + MLLT + SAT(tri3)

(Table-5): Results of speech recognition based on WER using language model matched to text of databases and language model weight equal one. Monophone(mono), triphone with delta and delta-delta (tri1), triphone with LDA+MLLT(ti2), triphone with LDA + MLLT + SAT(tri3), SGMM.

دادگان آزمون	مدل آوازی	WER
فارس دات	mono	16/7
	tri1	19/38
	tri2	22/4
	tri3	15/69
	SGMM	13/95
فارس دات	mono	17/61
	tri1	11/49
	tri2	10/94
	tri3	8/50
	SGMM	6/53

کلیدواژه انتخاب شد. نکته‌ای دیگر حجم دادگان مورد آزمون است، معیار ارزیابی، ATWV، متناسب با حجم و طول دادگان آزمون است. ATWV بیانگر مجموع تمامی کلیدواژه‌ها و عبارت‌ها می‌باشد؛ بنابراین در دادگان آزمون با حجم کم اگر در تشخیص یک کلیدواژه با تعداد تکرار کم اشتباہ شود، تأثیر بهسزایی در کارایی دارد. این نکته در دادگان فارس‌داد کوچک و اخبار اتفاق افتاده است. تعداد کلیدواژه‌ها برای دادگان فارس‌داد کوچک ۱۶۸ کلیدواژه، فارس‌داد بزرگ ۱۰۲۵۸ و اخبار فارسی ۳۵۹۱ در نظر گرفته شد. در جدول (۷) نتایج تشخیص عبارت‌های گفتاری بر اساس سامانه بازشناسی گفتار طراحی شده در جدول (۴) و معیار ATWV آمده است. در این نتایج برای بازشناسی دادگان اخبار فارسی فقط از مدل فارس‌داد بزرگ استفاده شده است.

همان‌طور که در قسمت چالش‌ها گفته شد، خطای بازشناسی گفتار بر روی تشخیص عبارت‌ها تأثیر زیادی دارد. با مقایسه جدول نتایج بازشناسی گفتار (جدول ۴) و جدول نتایج سامانه‌ی تشخیص (جدول ۷) می‌توان دید که با کاهش خطای بازشناسی، نتایج بهتری برای تشخیص عبارت‌ها کسب شده است. همچنین از روی نتایج گزارش شده قبلی [2] می‌توان گفت با خطای بازشناسی در حد بیست تا سی درصد، می‌توان به نتایج بالا و قابل قبولی در بخش تشخیص رسید که این هم از روی جدول‌ها به‌وضوح قابل مشاهده است.

(جدول-۷): نتایج تشخیص عبارت‌ها با معیار ATWV و MTWV برای کلیدواژه‌های INV. از سامانه‌های بازشناسی آموزش داده شده در جدول (۴) برای ساخت سامانه‌ی تشخیص استفاده شده است. (در دادگان اخبار فارسی از مدل آوازی فارس‌داد بزرگ استفاده شده و از مدل آوازی فارس‌داد کوچک به‌علت خطای بالا صرف نظر شده است).

(Table-7): Results of STD based on ATWV & MTWV for INV keywords. ASR in Table 4 used for STD. (In Persian news dataset we used only acoustic model based large FarsDat and regardless of small FarsDat since WER is too big).

MTWV	ATWV	مدل آوازی	دادگان
0/8456	0/8353	mono	فارس‌داد
0/7886	0/7814	tri1	
0/7731	0/7297	tri2	
0/7925	0/7638	tri3	
0/8418	0/8367	SGMM	
0/8949	0/8944	mono	فارس‌داد
0/9064	0/9063	tri1	
0/9078	0/9076	tri2	
0/9104	0/9098	tri3	
0/9206	0/9204	SGMM	بزرگ
0/4597	0/4378	mono	
0/2651	0/2387	tri2	
0/5795	0/5628	tri3	
0/8008	0/7933	SGMM	اخبار فارسی

است. دادگان آموزش فقط خوانش متون معیار فارسی است ولی دادگان اخبار فارسی مخلوطی از خوانش اخبار فارسی، گزارش‌ها در محیط‌های مختلف و گفتار همراه با موسیقی است.

برای ارزیابی بهتر سامانه بازشناسی اخبار فارسی از پیکره متنی ایستا شامل ۵۵۰ میلیون واژه برای ساخت مدل زبانی استفاده و نتایج در جدول (۶) نمایش داده شده است. با مقایسه نتایج بازشناسی اخبار فارسی با مدل‌های زبانی منطبق بر خود دادگان (جدول ۴) و مدل زبانی وسیع‌تر (جدول ۶) می‌توان تأثیر مدل زبانی را بهتر مشاهده کرد. همان‌طور که انتظار می‌رفت نتایج برای مدل زبانی منطبق بر دادگان بهتر است.

(جدول-۶): نتایج خطای بازشناسی گفتار دادگان اخبار فارسی بر مبنای معیار WER با مدل زبانی بزرگ‌تر از متن دادگان.

(Table-6): Results of speech recognition based on WER for Persian IRIB News with larger language model than text of database.

WER	وزن مدل زبانی	مدل آوازی	مدل زبانی	دادگان
78/80	11	mono	فارس‌داد	اخبار فارسی
83/25	12	tri1		
81/32	10	tri2		
60/22	15	tri3		
43/74	9	SGMM		

۶-۴- نتایج سامانه‌ی تشخیص عبارت‌های گفتاری همان‌طور که بحث شد، سامانه‌های تشخیص عبارت‌های گفتاری بر پایه بازشناسی گفتار داخل واژگان (INV) را می‌توانند تشخیص دهند و از آن جهت که واژه‌های خارج از واژگان (OOV) در خروجی بازشناسی گفتار هیچ اطلاعاتی از آن‌ها نیست، باید از روش‌های دیگر آن‌ها را تشخیص داد. در زیرتشخیص عبارات گفتاری از دو جهت INV و OOV ارزیابی شده است.

سامانه‌ی تشخیص واژگان داخل واژگان از روی الگوریتم معرفی شده در [5] ساخته شد. این الگوریتم در کلیدی [48] پایه‌ریزی شده است و با استفاده از آن، این سامانه طراحی شد.

۶-۵- کلیدواژه‌های درون واژگان سامانه بازشناسی گفتار (INV)

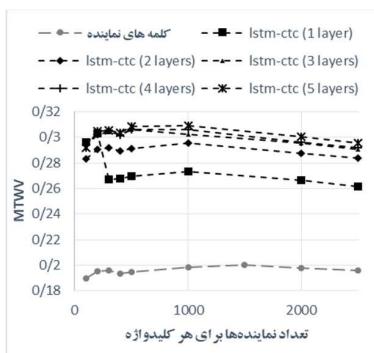
یکی از نکات در ارزیابی سامانه‌ی تشخیص عبارت‌های گفتاری، کلیدواژه‌های مورد جستجو در این سامانه‌ها است. از آنجایی که کلیدواژه‌های معیاری برای زبان فارسی وجود ندارد، تمامی واژه‌های یکتای پنج‌واحی دادگان به عنوان



واژه‌های نماینده و مرحله بعد خروجی‌های واژه‌های نماینده به یک سامانه بازناسی LSTM-CTC داده می‌شود، تا دوباره امتیازبندی و ارزیابی شوند. همان‌طور که اشاره شد، شبکه LSTM-CTC به صورت واژه به واژه و نه بر اساس گفتار پیوسته آموزش داده شده است و تنها خروجی‌های نماینده را بر اساس واژه‌های جداگانه بازناسی می‌کند و امتیاز اطمینان دوباره تخمین زده می‌شود.

در روش واژه‌های نماینده تعداد واژه‌های مشابه آوایی کلیدواژه مهم هستند و با افزایش تعداد واژه‌های نماینده بهازای کلیدواژه به طور معمول ΔTWV روند افزایشی و سپس کاهشی دارد [31]: این به خاطر افزایش هشدارهای نادرست است؛ هرچند که تعداد ازدست‌رفته‌ها کاهش می‌یابند. از آنجایی که خطای هشدار نادرست در ATWV دارای جرمیه دهبرابری نسبت به خطای ازدست‌رفته است، افزایش هشدارهای نادرست تأثیر بدتری بر روی ATWV می‌گذارد. تعداد نماینده‌ها بهازای هر کلیدواژه به تعداد واژه‌های واژگان بازناسی گفتار وابسته است. در اینجا به‌علت محدودی‌بودن تعداد واژه‌ها در واژگان (حدود ۳۸ هزار واژه) به نظر می‌رسد؛ تعداد نماینده‌ها بهازای هر کلیدواژه تأثیری بر ATWV نمی‌گذارد.

در شکل (۴) نتایج روش واژه‌های نماینده و روش بهبود آن، روش تصدیق LSTM-CTC بر مبنای معیار MTWV آمده است. با مقایسه نتایج، بهبود تا پنجاه درصد در این روش به دست آمده است. همان‌طور که در شکل‌های (۵) و (۶) مشاهده می‌شود، شبکه LSTM-CTC درصد خطای از دست‌رفته را کاهش می‌دهد ولی هم‌زمان خطای هشدار نادرست را افزایش می‌دهد ولی در مجموع معیار MTWV روند افزایشی نسبت به روش واژه‌های نماینده دارد، ولی نسبت به خود روش LSTM-CTC روند نزولی است.



(شکل-۴): نتایج حاصل از روش واژه‌های نماینده و روش بهبود آن، روش تصدیق LSTM-CTC بر مبنای معیار MTWV

(Figure-4): Results of proxy words and improve it based on LSTM-CTC validation.

در جدول (۸) زمان اجرای نمایه‌گذاری و جستجو بر روی داده‌های FarsDat بزرگ گزارش شده که بر روی سیستمی با مشخصات زیر اجرا شده است: Intel(R) Core(TM) i7-6500U CPU @ 2.50GHz, 2592 Mhz, 2 Cores, 4 Logical Processors

(جدول-۸): زمان اجرای نمایه‌گذاری و جستجو بر روی داده‌های FarsDat بزرگ

(Table-8): Time of indexing and searching for large FarsDat dataset

زمان (ثانیه)	پردازش
141	نمایه‌گذاری
77	جستجو

۴-۶-۲- کلیدواژه‌های خارج از واژگان سامانه بازناسی گفتار (OOV)

همان‌طور که در بخش قبل اشاره شد، حجم دادگان و تعداد کلیدواژه‌ها در معیار ارزیابی ATWV تأثیر بهسزایی دارد و در دادگان آزمون با مدت زمان کم اگر در تشخیص یک کلیدواژه با تعداد تکرار کم، استباهی رخ دهد، تأثیر بهسزایی در ATWV دارد. از آنجایی که حجم دادگان FarsDat کوچک کم است، ارزیابی کلیدواژه‌های OOV فقط بر روی دادگان FarsDat بزرگ انجام گرفت.

به‌علت اینکه تمامی واژه‌های دادگان درون واژگان سامانه‌های بازناسی گفتار طراحی شده بودند، تعداد ۵۱۵ کلیدواژه بین چهار واج تا هشت واج از واژگان بازناسی حذف شد و عمل بازناسی دوباره بر روی دادگان FarsDat بزرگ آموزش داده شد و مورد آزمون قرار گرفت. خطای WER این حالت از بازناسی گفتار در جدول (۹) آمده است.

(جدول-۹): نتایج مدل‌های بازناسی گفتار آموزش داده شده با حذف ۵۱۵ واژه از واژگان؛ برای بررسی روش‌های تشخیص کلیدواژه‌های OOV

(Table-9): Results of ASR with delete 515 words from the ASR lexicon, since evaluating STD for OOV keywords.

دادگان	مدل آوایی	وزن مدل زبانی	بهترین WER
فارسDat	mono	7	56/8456
	tri1	9	37/7886
	tri2	10	36/7731
بزرگ	tri3	12	36/7925
	SGMM	9	28/8418

سامانه تشخیص عبارت بر روی OOV بر پایه مشبك خروجی مدل آموزشی SGMM برای بازناسی و در دو مرحله طراحی شد. یک مرحله پایه‌ای بر اساس الگوریتم

حاضر پژوهش‌ها بر روی زبان‌هایی با داده‌های کم متتمرکز شده‌اند [29]. در اینجا بدليل وجود دادگان مختلف و حجم متفاوت آن‌ها نمی‌توان به معیار مشخصی برای مقایسه روش‌ها رسید.

در [4]، ده روش بر روی زبان اسپانیایی به کار گرفته شده است. همان‌طور که در مقاله عنوان شده است، دادگان MAVIR و EPIC دادگان معیار و با لهجه‌ی اسپانیایی و دادگان با تنوع دادگان بیشتری هستند و طبق معیار ATWV به ترتیب به اعداد ۰/۹۳ و ۰/۶۴ برای تشخیص عبارت‌های گفتاری رسیده‌اند. کاهش معیار بر روی این دادگان، بدلاً تفاوت آن‌ها محسوس است.

در مقاله‌های [6], [8] روش استفاده از بازشناسی گفتار بر زبان‌های با دادگان محدود بررسی شده است. برای مقایسه روش‌های فارسی، بهدلیل تفاوت دادگان و کلیدواژه‌ها نمی‌توان بررسی دقیقی انجام داد.

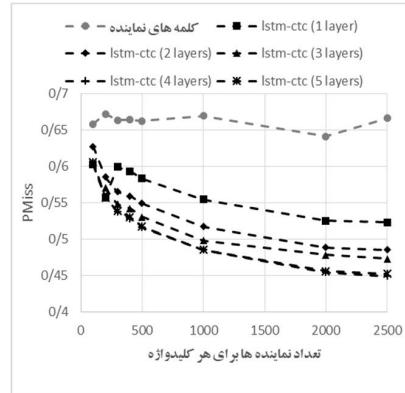
در [37] همانند این مقاله از دادگان فارس دات بزرگ استفاده شده و از روش مبدل عامل برای نمایه‌گذاری استفاده کرده و همچنین از دو نمایه‌گذاری واژه‌ای و واژی به صورت موازی استفاده کرده است. این پایان‌نامه به نزد خطای واژه و ATWV مشابه این مقاله رسیده ولی روشی برای جستجوی کلیدواژه‌های OOV بررسی نکرده است. در [9] که به نظر تکمیل‌کننده پایان‌نامه قبلی است، تمرکز بر روی دادگان محدود و کلیدواژه‌های OOV شده است. در این پایان‌نامه از روش واژه‌های نماینده استفاده شده و با روش گسترش واژه‌ها این روش را تقویت کرده است. با توجه به نزد بالای واژه‌های خارج از واژگان نتایج خوبی گرفته شده است. در [38] از روش جستجو تنها بر روی رشته واج نه واژه استفاده شده و از روش‌های مبدل عامل و فاصله‌یاب لونشتابین برای جستجو بهره گرفته است. در مقایسه با روش‌های دیگر در زبان‌های دیگر به نتایج عددی بسیار بزرگ‌تری رسیده است [4]¹ (حدود چهار برابر) ولی نحوه انتخاب کلیدواژه‌ها، نوع داده‌ها در این اعداد مهم است. مقاله کنونی در واقع نوعی تکمیل‌کننده روش‌های [9]، [37] با مقایسی بزرگ‌تر و با دادگان واقعی‌تر است.

۵- نتیجه گیری

در این مقاله یک سامانه تشخیص عبارت‌های گفتاری برای اخبار فارسی و دادگان فارس‌داد بزرگ ارائه شد. تشخیص

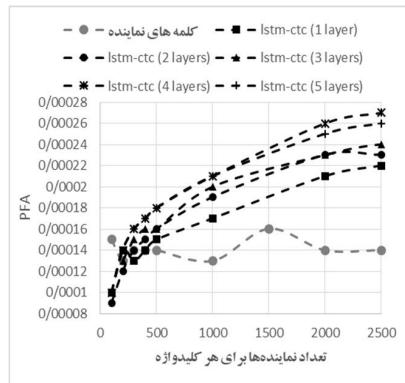
روش Phonc STD در مقاله مذکور.

در اینجا دوباره انتخاب کلیدواژه‌ها اهمیت پیدا می‌کنند، درواقع کلیدواژه‌های انتخابی برای کلیدواژه‌های OOV با واژه‌های داخل واژگان شباهت بالایی دارند. به طور مثال واژه «زیاد» به عنوان واژه خارج از واژگان انتخاب شده است ولی واژه‌های «زیاد و زیادی» در واژگان بازشناسی گفтар هستند که این می‌تواند عامل بالارفتن هشدارهای نادرست باشد. درنهایت با ارزیابی نتایج بر روی کلیدواژه‌های OOV و مقدار $ATWV$ برای معیار INV حاصل شد.



(شکل-۵): درصد خطای ازدست رفته حاصل از روش واژه‌های نماینده و روش بهبود آن، روش تصدیق LSTM-CTC.

(Figure-5): PMiss for proxy words and improve it based on LSTM-CTC validation.



(شکل-۶): در صد خطای هشدار نادرست حاصل از روش واژه‌های نماینده و روش بهبود آن، روش تصدیق LSTM-CTC.

(Figure-6): PFA for proxy words and improve it based on LSTM-CTC validation.

۴-۷- مقایسه با روش‌های پیشین فارسی

خلاصهایی از نتایج روش‌های مورداستفاده در زبان فارسی و دیگر زبان‌ها در جدول (۱) آمده است. در زبان انگلیسی بهدلیل وجود دادگان عظیم و متنوع چالش بازشناسی گفتار و به کارگیری آن در تشخیص عبارت‌های گفتاری دست کم در اخبار رسمی حل شده عنوان شده است [۵]. در حال



حجم پیکره متنی؛ به کارگیری از روش‌های نرم‌الیزه کردن نمره‌های عبارت‌های جستجو [58] برای نرمال‌کردن خروجی شبکه LSTM-CTC

6- References

- [1] L. Lee, J. Glass, H. Lee, and C. Chan, "Spoken Content Retrieval—Beyond Cascading Speech Recognition with Text Retrieval," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 9, pp. 1389–1420, Sep. 2015.
- [2] M. Larson and G. J. F. Jones, "Spoken Content Retrieval: A Survey of Techniques and Technologies," *Found. Trends® Inf. Retr.*, vol. 5, no. 3, pp. 235–422, 2012.
- [3] J. G. Fiscus, J. Ajot, J. S. Garofolo, and G. Doddington, "Results of the 2006 Spoken Term Detection Evaluation," *Proc. ACM SIGIR Work. Search. Spontaneous Conversational.*, pp. 51–55, 2006.
- [4] J. Tejedor *et al.*, "ALBAYZIN 2016 spoken term detection evaluation: an international open competitive evaluation in Spanish," *EURASIP J. Audio, Speech, Music Process.*, vol. 2017, no. 1, p. 22, 2017.
- [5] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC Spoken Document Retrieval Track: A Success Story," *Proc. TREC-8*, vol. 8940, no. 500–246, pp. 109–130, 1999.
- [6] J. Trmal *et al.*, "The Kaldi OpenKWS System: Improving Low Resource Keyword Search," *Interspeech2017*, pp. 3597–3601, 2017.
- [7] X. Anguera, L. J. Rodriguez-Fuentes, A. Buzo, F. Metze, I. Szoke, and M. Penagarikano, "QUESST2014: Evaluating Query-by-Example Speech Search in a zero-resource setting with real-life queries," *ICASSP. IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2015-Augus, pp. 5833–5837, 2015.
- [8] T. Alumäe *et al.*, "The 2016 BBN Georgian telephone speech keyword spotting system," *ICASSP. IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 5755–5759, 2017.
- [9] ز. گمار، "مدل‌های بیانی تمایزی برای تشخیص عبارات گفتاری در شرایط با منابع محدود،" پایان‌نامه کارشناسی ارشد، صنعتی شریف، ۱۳۹۴.
- [10] Z. Gomar, Discriminative Articulatory Models for Spoken Term Detection in Low-Resource Conditions, M.S. Thesis, Sharif University of Technology, 2016.
- [11] M. Crochemore, "Transducers and repetitions," *Theor. Comput. Sci.*, vol. 45, pp. 63–86, 1986.

۶- مراجع

عبارت‌های گفتاری یا جستجوی کلیدواژه‌ها روش‌هایی برای جستجوی دقیق یک عبارت و یا کلیدواژه درون یک سند گفتاری یا فایل صوتی ارائه می‌کنند. در این پژوهش از روشی بر پایه سامانه بازشناسی گفتار استفاده شد. در این روش ابتدا گفتار ورودی به شبکه بازشناسی می‌شود و سپس بر روی شبکه جستجو انجام می‌گیرد. یکی از چالش‌های این روش واژه‌های خارج از واژگان (OOV) است؛ این واژه‌ها در خروجی بازشناسی گفتار نمایش داده نمی‌شوند؛ بنابراین با روش‌های دیگری باید برای جستجوی آن‌ها را انجام داد. برای کلیدواژه‌های خارج از دادگان در این مقاله از روش واژه‌های نماینده کمک گرفته شد. در این روش به جای جستجوی مستقیم کلیدواژه، واژه‌هایی که از نظر آوایی شباهت به آن‌ها دارند و در بین واژگان ASR هستند، مورد جستجو قرار می‌گیرند.

در این پژوهش برای بازشناسی گفتار از مدل‌های مخفی مارکوف استفاده و از دادگان فارس‌داد بزرگ برای آموزش مدل آوایی استفاده شد. ارزیابی این مدل بر روی دادگان فارس‌داد بزرگ با روش آموزشی SGMM خطای ۲/۷۱ و بر روی دادگان اخبار فارسی خطای ۲۸/۲۳ حاصل شد. در تشخیص عبارت‌ها معیار ATWV با بازه بین منفی بی‌نهایت تا مثبت یک در نظر گرفته شد. تشخیص عبارت‌های گفتاری در دو سطح کلیدواژه‌های داخل واژگان (INV) و خارج واژگان (OOV) به طور جداگانه انجام گرفت. در سطح کلیدواژه‌های INV بر روی دادگان فارس‌داد بزرگ مقدار ۰/۹۲۰۶ و برای دادگان اخبار فارسی مقدار ۰/۸۰۰۸ نتیجه و در سطح کلیدواژه‌های OOV با استفاده از روش واژه‌های نماینده مقدار ۰/۲۰۰۲ برای دادگان فارس‌داد بزرگ حاصل شد. برای بهبود عملکرد این سامانه از یک شبکه LSTM-CTC بر پایه‌ی آموزشی واژه استفاده شد که درنهایت با حدود پنجاه درصد افزایش مقدار ۰/۳۰۵۸ حاصل شد.

برای ادامه پژوهش می‌توان به موارد زیر اشاره کرد: مطالعه بر روی انواع مختلف مدل‌کردن زیرواژه‌ها در زبان فارسی برای جستجوی کلیدواژه‌های خارج از دادگان؛ همانند [57] در زبان اسپانیایی و دیگر زبان‌ها [2]؛ استفاده از روش‌های بازیابی و روش‌هایی که در آن به صورت یکجا سامانه را نظر می‌گیرند [1]؛ استفاده از روش‌های آموزشی شبکه‌های عصبی برای مدل آوایی متناسب با مدت‌زمان دادگان؛ به کارگیری از مدل‌های زبانی با تعداد واژه‌های بیشتر و روش‌های آموزشی شبکه‌های عصبی متناسب با

- spoken term detection using attention-based multi-hop networks," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2018-April, pp. 6264–6268, 2018.
- [24] R. C. Rose and D. B. Paul, "A hidden Markov model based keyword recognition system," *Int. Conf. Acoust. Speech, Signal Process.*, pp. 129–132 vol.1, 1990.
- [25] J. G. Wilpon, L. R. Rabiner, C.-H. Lee, and E. R. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," *Ieee Taslp*, vol. 3, no. I, pp. 1870–1878, 1990.
- [26] A. Tavaneai, H. Sameti, and S. H. Mohammadi, "False alarm reduction by improved filler model and post-processing in speech keyword spotting," *IEEE Int. Work. Mach. Learn. Signal Process.*, 2011.
- [27] R. Sukkar and J. Wilpon, "A two pass classifier for utterance rejection in Keyword Spotting," *Acoust. Speech, Signal ...*, pp. 1–4, 1993.
- [28] M. G. Rahim, C. H. Lee, and B. H. Juang, "Discriminative utterance verification for connected digits recognition," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 3, pp. 266–277, 1997.
- [29] "KWS16 Evaluation Plan." [Online]. Available: <https://www.nist.gov/%0Asites/default/files/documents/itl/iad/mig/KWS16-evalplan-v04.pdf>.
- [30] C. Chelba, J. Silva, and A. Accro, "Soft indexing of speech content for search in spoken documents," *Comput. Speech Lang.*, vol. 21, no. 3, pp. 458–478, Jul. 2007.
- [31] G. Chen, O. Yilmaz, J. Trmal, D. Povey, and S. Khudanpur, "Using proxies for OOV keywords in the keyword search task," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013, pp. 416–421.
- [32] T. K. Chia, K. C. Sim, H. Li, and H. T. Ng, "Statistical lattice-based spoken document retrieval," *ACM Trans. Inf. Syst.*, vol. 28, no. 1, pp. 1–30, Jan. 2010.
- [33] Y. C. Pan and L. S. Lee, "Performance analysis for lattice-based speech indexing approaches using words and subword units," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 6, pp. 1562–1574, 2010.
- [34] W. Hartmann, V. B. Le, A. Messaoudi, L. Lamel, and J. L. Gauvain, "Comparing decoding strategies for subword-based keyword spotting in low-resourced languages," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, no. September, pp. 2764–2768, 2014.
- [35] L. S. Lee and Y. C. Pan, "Voice-based information retrieval - How far are we from the text-based information retrieval?," *Proc. 2009*
- [11] J. S. Bridle, "An efficient elastic-template method for detecting given words in running speech," *Brit. Acoust. Soc. Meet.*, pp. 1–4, 1973.
- [12] A. Mandal, K. R. Prasanna Kumar, and P. Mitra, "Recent developments in spoken term detection: a survey," *Int. J. Speech Technol.*, vol. 17, no. 2, pp. 183–198, Jun. 2014.
- [13] J. Bridle, "An efficient elastic template method for detecting given keywords in the running speech," *Proc. Br. Acoust. Soc. Meet.*, pp. 1–4, 1973.
- [14] C. Parada, A. Sethy, and B. Ramabhadran, "Query-by-example spoken term detection for OOV terms," *Proc. 2009 IEEE Work. Autom. Speech Recognit. Understanding, ASRU 2009*, pp. 404–409, 2009.
- [15] J. Tejedor, I. Szöke, and M. Fapso, "Novel methods for query selection and query combination in query-by-example spoken term detection," *Proc. 2010 Int. Work. Search. spontaneous conversational speech - SSCS '10*, pp. 15–20, 2010.
- [16] M. C. Madhavi and H. A. Patil, "Partial matching and search space reduction for QbE-STD," *Comput. Speech Lang.*, vol. 45, pp. 58–82, Sep. 2017.
- [17] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriograms," *Proc. 2009 IEEE Work. Autom. Speech Recognit. Understanding, ASRU 2009*, pp. 398–403, 2009.
- [18] M. Huijbregts, M. McLaren, and D. Van Leeuwen, "Unsupervised acoustic sub-word unit detection for query-by-example spoken term detection," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 4436–4439, 2011.
- [19] P. Fousek and H. Hermansky, "Towards ASR Based on Hierarchical Posterior-Based Keyword Recognition," *2006 IEEE Int. Conf. Acoust. Speech Signal Process. Proc.*, vol. 1, pp. I-433–I-436.
- [20] H. Sakoe and S. Shiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 26, no. 1, pp. 43–49, 1978.
- [21] C. Chan and L. Lee, "Unsupervised Spoken-Term Detection with Spoken Queries Using Segment-based Dynamic Time Warping," *Evaluation*, no. September, pp. 693–696, 2010.
- [22] D. Ram, L. Miculicich, and H. Bourlard, "CNN based query by example spoken term detection," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2018–September, pp. 92–96, 2018.
- [23] C. W. Ao and H. Y. Lee, "Query-by-example



- [42] H. Naderi, Keyword Spotting in Speech Utterance, M.S. thesis, Shahrood University of Technology, 2013.
- [43] M. Bijankhan, J. Sheikhzadegan, M. R. Roohani, Y. Samareh, C. Lucas, and M. Tebyani, "FARSDAT-The speech database of Farsi spoken language," *Proc. Aust. Conf. Speech Sci. Technol.*, vol. 2, no. 0, pp. 826–831, 1994.
- [44] J. Sheikhzadegan and M. Bijankhan, "Persian speech databases," *2nd Work. Persian Lang. Comput.*, pp. 247–261, 2006.
- [45] M. Bijankhan, J. Sheykhzadegan, M. R. Roohani, R. Zarrintarc, S. Z. Ghasemi, and M. E. Ghasedi, "Tfarsdat - The telephone farsi speech database," *EUROSPEECH 2003 - 8th Eur. Conf. Speech Commun. Technol.*, 2003.
- [46] D. Can and M. Sarclar, "Lattice Indexing for Spoken Term Detection," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 8, pp. 2338–2347, Nov. 2011.
- [47] Z. Lv, J. Kang, W. Q. Zhang, and J. Liu, "An LSTM-CTC based verification system for proxy-word based OOV keyword search," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 5655–5659, 2017.
- [48] C. Parada, A. Sethy, and B. Ramabhadran, "Balancing false alarms and hits in spoken term detection," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, pp. 5286–5289, 2010.
- [49] Z. Victor, S. Seneff, and J. Glass, "TIMIT acoustic-phonetic continuous speech corpus," *Speech Commun.*, vol. 9, no. 4, pp. 351–56, 1990.
- [50] ب. باباعلی, "پایه گذاری بسترهای نو و کارآمد در حوزه بازشناسی گفتار فارسی," پردازش علایم و داده ها, شماره ۱۳ (۳)، صص ۵۱-۶۲، ۱۳۹۵.
- [50] B. BabaAli, "State-of-the-art and Efficient Framework for Persian Speech Recognition, *jsdp* , Vol (3), pp. 51-62, 2017.
- [51] م. اسلامی، م. شریفی آتشگاه، ص. علیزاده لمجیری ط. زندی، "واژگان زیبای زبان فارسی," مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه، ۱۳۸۳.
- [51] M. Eslami, M. Sharifi Atashgah, S. Alizade , T, Zandi, "Persian Generative Lexicon" Proceedings of the first Persian language and computer research workshop, 2005.
- [52] " " Hazm." [Online]. Available: <https://github.com/sobhc/hazm>.
- [36] D. Can, "Indexation, retrieval & decision techniques for spoken term detection," PhD diss, Boğaziçi University, 2010.
- [37] M. Qadiri Nia, "Design and Performance Improvement of a Spoken Term Detection System", M.S. thesis, Sharif Universitt of Technology, 2015.
- [38] M. Abbassian, "Keword Spotting in Persian Speech Using a Hybrid Model of DNN and HMM", M.S. thesis, Amir Kabir University of Technology, 2017.
- [39] S.S. Sarfjou, Introducing a New Information Retrieval Framework for Persian Speech Retrieval, M.S. thesis, Qom University, 2012.
- [40] م. ی. اخلاقی, "ارائه یک روش جدید بازیابی اطلاعات مناسب برای متون حاصل از بازشناسی گفتار," پایان نامه کارشناسی ارشد، دانشگاه قم، ۱۳۹۲.
- [41] M.Y. Akhlaqi, "Introducing a New Information Retrieval Method for Speech Recognized Texts, M.S. thesis, Qom University, 2014.
- [42] م. ح. سلطانی، "چارچوبی جدید برای بازشناسی گفتار به منظور استفاده در بازیابی صدای گفتاری," پایان نامه کارشناسی ارشد، دانشگاه قم، ۱۳۹۲.
- [43] M.H. Soltani, "Introducing a New Information Retrieval Framework for Speech Retrieval" M.s. thesis, Qom University, 2014.
- [44] ه. نادری، "جستجوی کلمات کلیدی در رشته‌ی گفتار," پایان نامه کارشناسی ارشد، دانشگاه صنعتی شاهرود، ۱۳۹۱.
- [37] م. غدیری‌نیا, "طراحی و بهبود یک سامانه‌ی تشخیص اصطلاحات گفتاری," پایان نامه کارشناسی ارشد، صنعتی شریف، ۱۳۹۳.
- [38] م. عباسیان, "شناسایی کلمات کلیدی در گفتار فارسی توسط سیستم تلفیقی مدل مخفی مارکوف و شبکه‌های عصبی عمیق," پایان نامه کارشناسی ارشد، دانشگاه صنعتی امیرکبیر، ۱۳۹۵.
- [39] س. س. صرفجو, "چارچوبی جدید برای بازیابی اطلاعات به منظور استفاده در بازیابی صدای گفتاری فارسی," پایان نامه کارشناسی ارشد، دانشگاه قم، ۱۳۹۰.
- [40] م. ی. اخلاقی, "ارائه یک روش جدید بازیابی اطلاعات مناسب برای متون حاصل از بازشناسی گفتار," پایان نامه کارشناسی ارشد، دانشگاه قم، ۱۳۹۲.
- [41] م. ح. سلطانی، "چارچوبی جدید برای بازشناسی گفتار به منظور استفاده در بازیابی صدای گفتاری," پایان نامه کارشناسی ارشد، دانشگاه قم، ۱۳۹۲.
- [42] ه. نادری، "جستجوی کلمات کلیدی در رشته‌ی گفتار," پایان نامه کارشناسی ارشد، دانشگاه صنعتی شاهرود، ۱۳۹۱.





اعظم باستان فرد, درجه دکترا را در رشته علوم رایانه از دانشگاه صنعتی توکیو و فوق دکترا را در سال ۱۳۸۴ در گرافیک رایانه‌ای از دانشگاه ژنو کسب کرده‌اند. ایشان هم‌اکنون به عنوان عضو هیأت علمی دانشگاه آزاد واحد کرج مشغول به تدریس دروس نرم‌افزار و پردازش داده است. زمینه‌های پژوهشی ایشان پردازش صدا و تصویر، یادگیری ماشین، گرافیک رایانه‌ای و بازی‌های رایانه‌ای است.

نشانی رایانمه ایشان عبارت است از:
bastanfard@kiau.ac.ir

- [53] M. Federico, N. Bertoldi, and M. Cettolo, “IRSTLM: An open source toolkit for handling large scale language models,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 1618–1621, 2008.
- [54] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, “The DET Curve in Assessment of Detection Task Performance,” *Proc. Eurospeech '97*, pp. 1895–1898, 1997.
- [55] D. Jurafsky and J. H. Martin, *Speech and language processing*. 1999.
- [56] D. Povey *et al.*, “The subspace Gaussian mixture model - A structured model for speech recognition,” *Comput. Speech Lang.*, vol. 25, no. 2, pp. 404–439, 2011.
- [57] J. Tejedor, D. Wang, J. Frankel, S. King, and J. Colás, “A comparison of grapheme and phoneme-based units for Spanish spoken term detection,” *Speech Commun.*, vol. 50, no. 11–12, pp. 980–991, 2008.
- [58] Y. Wang and F. Metze, “An in-depth comparison of keyword specific thresholding and sum-to-one score normalization,” *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 2474–2478, 2014.



هادی ویسی مدرس دکترای خود را در رشته مهندسی رایانه در سال ۱۳۹۰ از دانشگاه صنعتی شریف دریافت کرده و از سال ۱۳۹۱ عضو هیأت علمی دانشگاه تهران است. زمینه‌های تخصصی ایشان پردازش زبان طبیعی و پردازش گفتار، یادگیری ماشین، شبکه‌های عصبی مصنوعی و یادگیری عمیق است.

نشانی رایانمه ایشان عبارت است از:
h.veisi@ut.ac.ir



سید اکبر قریشی مدرس کارشناسی خود را در سال ۱۳۹۳ در رشته مهندسی برق (گرایش مخابرات) از دانشگاه یزد دریافت کرد. ایشان در سال ۱۳۹۸ از دانشگاه صدا و سیما در مقطع کارشناسی ارشد در رشته مهندسی صدا فارغ‌التحصیل شد. وی از سال ۱۳۹۷ عضو آزمایشگاه پردازش علائم و داده‌های دانشگاه تهران است. زمینه مورد علاقه وی پردازش گفتار و بازشناسی گفتار است.

نشانی رایانمه ایشان عبارت است از:
akbar20gh@gmail.com

فصلنامه
پژوهش‌های
دانشگاه
پردازش گفتار

