ՀՀ ԳԱԱ ԻՆՖՈՐՄԱՏԻԿԱՅԻ ԵՎ ԱՎՏՈՄԱՏԱՑՄԱՆ ՊՐՈԲԼԵՄՆԵՐԻ ԻՆՍՏԻՏՈՒՏ

Դավիթ Սպարտակի Քարամյան

Խոսքի հուսալի մշակում՝ ներդրված արհեստական բանականության կիրառություններում

Ե.13.05 – «Մաթեմատիկական մոդելավորում, թվային մեթոդներ և ծրագրերի համալիրներ» մասնագիտությամբ տեխնիկական գիտությունների թեկնածուի գիտական աստիճանի հայցման ատենախոսության

ՍԵՂՄԱԳԻՐ

ԵՐԵՎԱՆ – 2024

INSTITUTE FOR INFORMATICS AND AUTOMATION PROBLEMS OF THE NAS RA

Davit Spartak Karamyan

Robust speech processing in embedded AI applications

SYNOPSIS

of the dissertation for obtaining a Ph.D. degree in Technical Sciences on specialty 05.13.05 ”Mathematical modeling, digital methods and program complexes”

YEREVAN - 2024

Ատենախոսության թեման հաստատվել է Հայ Ռուսական համալսարանում

Գիտական ղեկավար՝	Ֆիզ. մաթ. գիտ. թեկնածու Ա.Ն. Հարությունյան
Պաշտոնական ընդդիմախոսներ՝	Ֆիզ. մաթ. գիտ. xxxxx X. xxxxxxxxxx
	Ֆիզ. մաթ. գիտ. xxxxx X.X. xxxxxxxxxx
Առաջատար կազմակերպություն՝	XX XXX xxxxxxxxxx xxxxxxxxxx

Ատենախոսության պաշտպանությունը կկայանա 2024թ. xxxxxx X-ին, ժ. XX:YY-ին ԳԱԱ Ինֆորմատիկայի և ավտոմատացման պրոբլեմների ինստիտուտում գործող XXX մասնագիտական խորհրդի նիստում հետևյալ հասցեով՝ Երևան, 0014, Պ. Սևակի 1:

Ատենախոսությանը կարելի է ծանոթանալ ՀՀ ԳԱԱ ԻԱՊԻ գրադարանում:

Սեղմագիրն առաքված է 2024թ. xxxx yy-ին:

Մասնագիտական խորհրդի գիտական
քարտուղար ֆիզ. մաթ. գիտ. դոկտոր՝ Մ. Ե. Հարությունյան

The topic of the dissertation was approved at the Russian Armenian University

Scientific supervisor:	Ph.D. of Phys. Math. Sciences A.N. Harutyunyan
Official opponents:	XX XXX xxxxxxxxxx xxxxxxxxxx
	XX XXX xxxxxxxxxx xxxxxxxxxx
Leading organization:	XX XXX xxxxxxxxxx xxxxxxxxxx

The dissertation defence will take place on XXXXX YY, 2024; at XX:YY, at the Specialized Council XXX «Informatics» at the Institute of Informatics and Automation Problems of NAS RA.Address: Yerevan, 0014, P. Sevak 1.

The dissertation is available in the library of IIAP NAS RA.

The abstract is delivered on xxx yy, 2024.

Scientific Secretary of the Specialized Council, D.Ph.M.S. M. E. Haroutunian

Relevance of the theme

Speech recognition systems have advanced significantly in the past decade. Still, even with these remarkable advances, machines have difficulties understanding natural conversations with multiple speakers, such as in broadcast interviews, meetings, telephone calls, videos or medical recordings. One of the first steps in understanding natural conversations is to recognize the spoken words and their corresponding speakers. **Speaker Diarization (SD)** determines "who spoke when" in multi-speaker audio and is a crucial part of the speech translation system. SD is used in conjunction with **Automatic Speech Recognition (ASR)** to assign a speaker label to each transcribed word and has widespread applications in generating meeting/interview transcripts, medical notes, automated subtitling and dubbing, downstream speaker analytics, among others (we refer to this combined system as **ASR-SD** in this thesis). Usually, this is done in multiple steps that include (1) transcribing the words using an ASR system, (2) predicting "who spoke when" using a speaker diarization system, and, finally, (3) reconciling the output of those two systems [1]. A typical reconciliation algorithm works as follows: (1) If the word segment overlaps with at least one speaker segment, then this word is associated with the speaker that has the biggest temporal overlap with this word; (2) otherwise, if this word segment does not overlap with any speaker segment, then it is associated with the speaker that has the smallest temporal distance to this word based on the segment boundaries [2]. This reconciliation algorithm is illustrated in Figure 1.

ASR is a challenging task in **Natural Language Processing (NLP)**. It consists of multiple subtasks, including speech segmentation, acoustic and language modelling. Luckily, the introduction of **Connectionist Temporal Classification (CTC)** [3] removed the need for pre-segmented data and allowed the **Deep Neural Network (DNN)** to be trained end-to-end directly for sequence labelling tasks like ASR. As a result, a CTC-based ASR pipeline consists of the following blocks (as shown in Figure 2):

1. A **Voice Activity Detection (VAD)** to exclude non-speech segments.
2. **Feature Extraction:** Audio signal pre-processing using normalization, windowing, spectrogram (typical spectrogram types are **Mel-spectrogram** and **Mel Frequency Cepstral Coefficients (MFCC)** [4, 5]). The spectrograms are used to convert audio signals into a form that the system can process. They display how the intensity of different frequencies varies over time, providing a detailed picture of the sound.
3. **Acoustic Model:** A Deep Neural Network [6–10] that predicts the probability distributions $P(c, t)$ over vocabulary characters (or tokens) c per each time step t.
4. **Decoder:**
   (a) **Greedy Decoding:** Is the most straightforward strategy for a decoder. The letter with the highest probability is chosen at each timestep without regard to any semantic understanding of what is being said. Then, the repeated characters are collapsed, and *blank* tokens are discarded.
   (b) A **Language Model (LM)** [10, 11] can be used to add context and correct errors in the acoustic model. The LM is used in fusion with beam search decoding to find the best translation candidates. During the decoding phase, multiple alternative token sequences or *beams* are generated and then scored using the acoustic and language models to select the most likely translation sequence. A beam search decoder tries to determine what was spoken by combining what the acoustic model thinks it heard with the likely next word.
5. **Output:** The ASR model outputs words with corresponding timestamps and confidence scores.

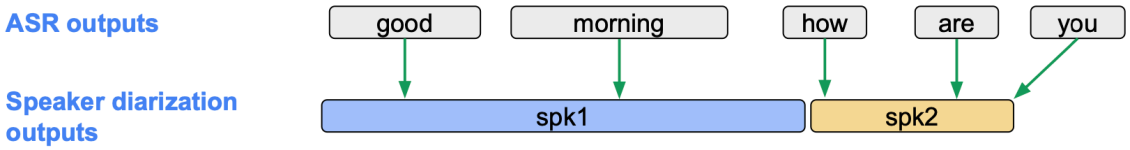

Figure 1: The reconciliation module associates each word from the ASR transcript with a speaker label from the speaker diarization outputs. The words ”good”, ”morning”, ”are” and ”you” are associated with the only speaker label that overlaps with them. The word ”how” overlaps with both spk1 and spk2 but has bigger overlaps with spk2; thus, it is associated with spk2. The word ”you” does not overlap with any speaker but is closest to spk2; thus, it is associated with spk2 [2].

Speaker diarization is a process to answer ”who spoke when” by segmenting audio recordings by speaker labels. To determine ”who spoke when”, SD systems need to capture the voice characteristics of unseen speakers and tell apart which areas in the audio signal belong to which speaker. To achieve this, speaker diarization systems extract voice characteristics, estimate the number of speakers, and then assign the audio segments to the corresponding speaker label. A typical SD system usually consists of several steps, as illustrated in Figure 2:

1. **Voice Activity Detection:** Identifying speech segments within the audio.
2. **Segmentation:** Generate speaker segments from the speech segments using a uniform sliding window segmentation or detecting speaker turns.
3. **Speaker Embedding Extraction:** Extract speaker embeddings (voice characteristics) for each speaker segment [12–16].
4. **Clustering:** Grouping speaker embeddings into clusters using algorithms such as **Spectral Clustering (SC)** [17–19] or **Agglomerative Hierarchical Clustering (AHC)** [20].
5. **Output:** The result is speech segments labelled with corresponding speaker identities.

Challenges in speech recognition

What makes ASR a difficult problem?

- **Open Vocabulary:** The complexity of ASR tasks varies with the size of the vocabulary involved. For instance, tasks with limited vocabularies, such as distinguishing between *yes* and *no* or recognizing numbers (*zero* to *nine*) in **digit recognition**, can achieve high accuracy. However, tasks like video transcription or capturing human dialogues, which may involve large vocabularies of up to 60,000 words, are significantly more challenging.
- **Conversational speech:** Another key variable is the nature of the speaker’s interaction. Speech directed towards machines, like dictation or interactions with a dialogue system, is typically easier to process than speech between humans. **Read speech**, such as that found in audiobooks where individuals read aloud, is also simpler to recognize. The most challenging scenario is **conversational speech**, like transcribing discussions in a business meeting, where two humans are conversing.
- **Acoustics:** The environment (or *channel*) and **background noise** also play a crucial role in ASR. Noise refers to any unwanted sound that is not part of the target speech signal. Unlike the human auditory system, which can adapt and filter out irrelevant sounds to focus on speech, ASR systems

struggle to distinguish between speech and noise, especially when the noise levels are high or the noise characteristics closely mimic those of speech. Speech captured in quiet settings with close-range microphones is more readily recognized than speech recorded in noisy environments, like busy streets or inside a car with an open window. Noise can be broadly categorized into several types:

  - *Ambient noise:* This includes sounds from the environment, such as traffic, wind, rain, or the hum of machinery and electronics. These sounds can be relatively constant or vary significantly over time.
  - *Reverberation:* Sound reflections from walls, floors, and ceilings can cause reverberation, making speech sound distant or echo. This is particularly challenging in large or sparsely furnished rooms.
  - *Transient noise:* Short, abrupt sounds like door slams, keyboard typing, or phone notifications are considered transient noise. They can momentarily obscure speech sounds, making it difficult for ASR systems to capture speech accurately.
  - *Competing speech:* While not fitting the constraint of being "non-speech," competing speech or overlapping conversations can severely impact ASR performance.

  For the context of this thesis, we will assume that ***background noise*** *does not encompass any competing speech sounds. Beyond this, there are no specific constraints on the noise distribution: it can be random or have specific patterns, and its intensity can vary widely.*

- **Speaker characteristics: Accent** and **speaker characteristics** significantly influence recognition ease. ASR systems perform best with speakers who use dialects or speech varieties similar to those used in their training. Speech from speakers with regional or ethnic accents, children, or those with unique pronunciations, intonations, and speaking speeds can be challenging to recognize if the system is only trained on speakers of standard dialects or only adult speakers [21].

Challenges in speaker diarization

Despite recent advancements in speaker diarization [17, 22], several factors make solving SD task difficult:

- **Uniform speaker segmentation:** Long speech segments very likely contain speaker turn boundaries, while short speech segments carry insufficient speaker information. This balance is crucial because overly long segments might miss the precise moments of change, whereas too short segments might not provide enough data for accurate speaker identification, leading to increased complexity in the diarization process.
- **Unknown number of speakers:** In general, both the identity of the speakers and the number of speakers are unknown beforehand. Estimating the number of speakers can introduce errors, especially in complex auditory environments or when speakers have similar voice characteristics, complicating the task of accurately separating and identifying individual speakers.
- **Speaker talk time:** A speaker needs to talk long enough to be accurately detected. If the talk time of a speaker is short, there's a significant risk that their speech might be incorrectly assigned to a more dominant speaker who has spoken for longer periods within the same audio segment.
- **Overlap speech:** Talking over each other or interrupting.

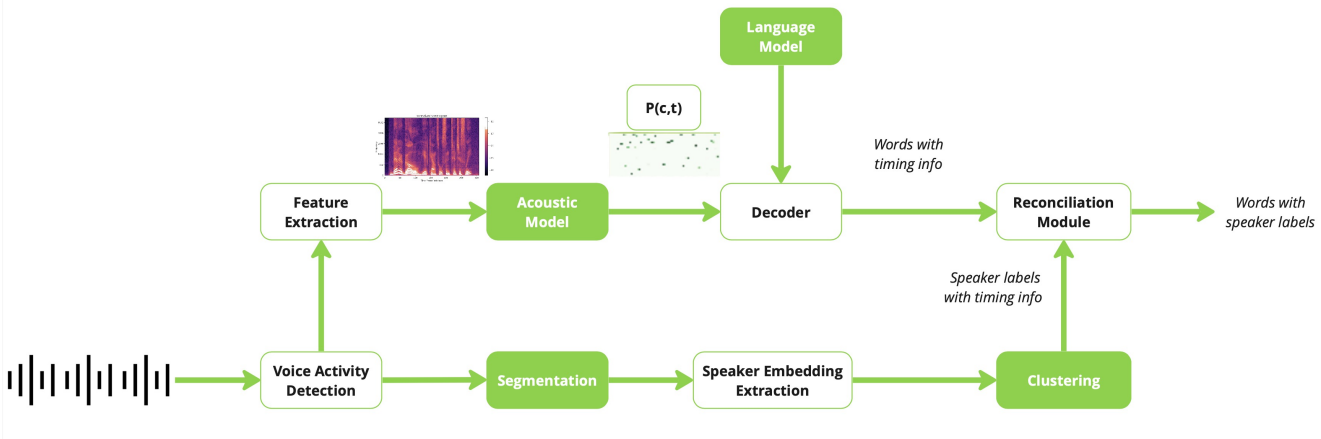

Figure 2: Schematic architecture for an ASR-SD system.

- **Acoustics:** Environmental sounds and room conditions can interfere with speaker recognition.
- **Consisting of multiple steps:** The SD system involves several steps (VAD → segmentation → embedding extraction → clustering), each of which introduces some amount of error.

ASR-SD for embedded application

In embedded applications, audio processing time (latency) and memory footprint are as important as quality. As always, trade-offs exist between processing time, memory space and accuracy [23]. Most current state-of-the-art speech recognition and speaker diarization systems are cloud-based. This thesis mainly focuses on systems that function on devices with limited resources (CPU and RAM), which require fast processing speeds and minimal memory footprints while maintaining high-quality output.

Running ASR-SD system directly into devices offers several practical advantages over cloud-based solutions, including reduced latency, enhanced privacy, and a decrease in reliance on internet connectivity:

- **Latency:** By processing audio data directly on the device, embedded ASR-SD systems significantly reduce the time between speech input and text output. This is particularly beneficial in applications requiring real-time feedback, such as interactive learning tools or instant transcription services.
- **Privacy:** Embedded systems process data locally, minimizing the risk of exposing sensitive information during transmission to and from cloud servers. This is especially important in fields like healthcare and legal services, where confidentiality is paramount.
- **Bot-Free:** Embedded ASR-SD systems ensure that the transcription and speaker diarization are conducted within the device without needing external bots or internet-based services. This setup enhances privacy and ensures that meetings can be conducted without interruptions or dependencies on external services. In environments where confidentiality and security are critical, such as boardroom meetings or sensitive negotiations, a bot-free environment is essential to prevent potential eavesdropping or data breaches.

The aim of the work

The techniques presented in this thesis address the robustness issues and improve the performance of current state-of-the-art speech recognition and speaker diarization systems. The thesis proposes strategies to address five challenges, each targeting different components of the ASR-SD system, which we refer to as robust ASR-SD (**rASR-SD**). The five identified subjects are:

1. Improving noise robustness of speech recognizer.
2. Robust keyword biasing methods to enhance the ASR system's ability to recognize terms and keywords without modifying the ASR model.
3. Robust speaker embedding extraction to enhance speaker recognition and diarization accuracy in noisy and reverberant environments.
4. Speaker error correction module using lexical information, leveraging the power of modern language models.
5. Various methods to accelerate the deployment of ASR-SD system on resource-constrained devices.

The practical significance of the work

ASR and SD technologies have revolutionized how we interact with audio data, enabling the automatic transcription of spoken words and the identification of individual speakers within a conversation. These technologies have found applications in a variety of fields, including but not limited to:

- **Meeting Assistant:** ASR and SD facilitate automatic meeting notes generation, making documenting discussions and decisions made during meetings easier. This application ensures that no critical information is lost and enhances meeting efficiency by providing searchable transcripts.
- **Medical Records:** In the healthcare sector, distinguishing between doctor and patient speech allows for the accurate transcription of medical consultations. This aids in the creation of more precise and detailed medical records, improving patient care and documentation.
- **Education:** In educational settings, the ASR-SD system can differentiate between teacher and student dialogues. This technology can be used to create lecture transcripts, enhance accessibility for students with disabilities, and support remote learning environments.
- **Call Centers:** For customer service interactions, identifying client versus customer speech is crucial. It enables the analysis of customer service calls for quality assurance, training purposes, and improving customer satisfaction.
- **Interviews:** In journalistic or research interviews, distinguishing between interviewer and interviewee ensures that the transcription accurately reflects the flow of conversation, aiding in the analysis and reporting of the interview content.

Integrations

The developed system has been successfully deployed at "Krisp.ai" LLC. It is currently being employed in the "Krisp AI Meeting Assistant" tool[1], for automatically documenting discussions and decisions made during online meetings.

[1] https://krisp.ai/ai-meeting-assistant/

The methods of investigation

In this thesis, we have used a wide range of approaches from different fields, including signal processing, machine learning, deep learning, probability theory, linear programming, optimization theory, finite automata theory and related fields. The Python and C++ programming languages and their related packages were used to train deep neural networks, process data and design algorithms. Previous related results also served as a basis for this work.

Approbation of the results

The obtained results were reported in several international and local scientific conferences and workshops:

1. D. Karamyan et al., "The Krisp Diarization system for the VoxCeleb Speaker Recognition Challenge 2023", The VoxCeleb Speaker Recognition Challenge 2023 (VoxSRC-23), Interspeech 2023 workshop, Dublin, Ireland, August 20-24, 2023.

2. D. Karamyan, "Multilingual Speaker Recognition Benchmark", Science and Technology Convergence Conference (STCC), Yerevan, Armenia, September 28-29, 2022.

3. D. Karamyan et al., "Compact N-Gram Models for Armenian", Collaborative Technologies and Data Science in Smart City Applications (CODASSCA 2022), Yerevan, Armenia, August 23-26, 2022.

4. D. Karamyan, "Adaptive Noise Cancellation for Robust Speech Recognition in Noisy Environments", Reporting Conference-2024, Yerevan State University, Yerevan, Armenia, April 10-15, 2024.

The results were presented during scientific seminars at Russian-Armenian and Yerevan State University. Our proposed speaker diarization system emerged as the winner in the Voxceleb Speaker Recognition Challenge (VoxSRC) 2023[2], a widely recognized competition for evaluating speaker diarization systems [37].

Publications

All results are new and are published in local journals. The main results of this thesis have been published in 7 scientific articles in journals. The list of the articles is given at the end of the Synopsis.

Structure and scope of work

The dissertation consists of 5 chapters and a list of used literature. The thesis is written in X pages and has Y literature references. The thesis contains 31 figures and 18 tables.

The thesis is organized as follows:

- *Chapter 1* serves as an introduction. It describes the problem, the main challenges in speech recognition and speaker diarization, and the aim of the thesis.

- *Chapter 2* summarizes the fundamentals and background information on speech recognition and speaker diarization. It covers each topic in detail, along with the state-of-the-art work related to each area.

[2] https://mmai.io/datasets/voxceleb/voxsrc/interspeech2023.html

- *Section 3.1* introduces a mathematical framework for modelling noise and reverberation.
- *Section 3.2* introduces a new method called Weakly Noise Cancellation, proposed to address the challenge of integrating Noise Cancellation with the ASR model to enhance the robustness of the speech recognizer in noisy environments.
- *Section 3.3* tackles an important issue in speech recognition: the ASR system's ability to recognize unknown or rare words, such as technical terms and names. This section first analyzes the effectiveness of widely used keyword biasing techniques, then proposes a combined solution, and concludes with comparative studies of keyword biasing methods.
- *Section 3.4* explores several strategies to improve the accuracy of speaker recognition and diarization in noisy and reverberant environments, including multi-condition training, teacher-student learning and consistency regularization techniques.
- *Section 3.5* discusses the main challenges of combining ASR and SD systems and presents an error analysis of the ASR-SD system at the word level. It introduces two novel speaker error correction methods based on lexical information to correct errors around speaker turns.
- *Chapter 4* describes various methods to accelerate ASR-SD deployment on resource-constrained devices, including spectrogram downsampling, model quantization, subword language modelling and two-stage clustering techniques.
- Finally, *Chapter 5* concludes the thesis with a summary of the contributions made.

The main results of the work

The following points summarize the key contributions and findings:

1. **Improving noise robustness of speech recognizer:** We address the challenges faced when combining Noise Cancellation (NC) [24, 25] and speech recognition models. Applying noise cancellation before speech recognition may negatively impact word recognition accuracy, even if the audio remains audible to humans. This is because noise cancellation eliminates noisy segments forcefully, causing a shift in the input data distribution that ultimately affects speech recognition (see Figure 3).

   To address this issue, we propose a new method called Weakly Noise Cancellation (WNC) that softens the effect of noise reduction without requiring retraining ASR model. The proposed method was able to improve the accuracy of speech recognition compared to the baseline ASR model obtained with multi-condition training [26].

   Additionally, we found that augmenting the training process with noise cancellation further improves word recognition accuracy. For example, when there is 0 dB of noise, meaning that the noise level is the same as the actual voice, we aim to decrease the Word Error Rate (WER) by 1.2% compared to the baseline model trained with noise augmentations.

2. **Keyword biasing methods:** Most ASR systems face difficulties recognizing unknown or uncommon words such as person names, location names or technical terminologies that are rarely or never seen during training. The recently introduced *Whisper Large* [8] model is proficient at recognizing terms, names and other commonly used keywords due to its training on vast amounts of data. Nevertheless, using this model in **embedded AI** applications that require fast inference and have limited memory resources may not be practical. Aside from the computational aspect,

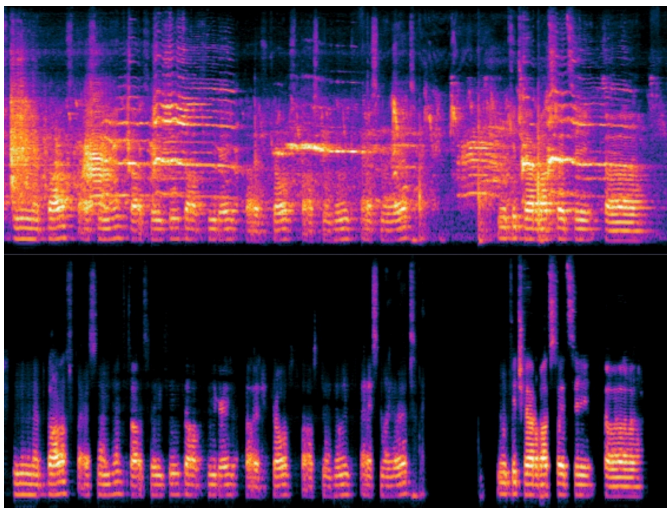

Figure 3: The top image represents the audio spectrogram before noise cancellation and the bottom one displays the audio spectrogram after noise cancellation.

it cannot recognize novel words that were not part of the training set. This problem is not only limited to ASR technology, but also applies to humans, as it is difficult to understand a conversation full of unknown words. These words often play a significant role in understanding the overall conversation despite their low frequency of occurrence.

Keyword biasing (also known as *contextual ASR* or *contextual biasing*) is a family of methods of guiding an ASR system towards a specified list of keywords and phrases provided along with the audio to be transcribed. We conduct a comparative analysis of various decoding strategies, including *CTC prefix beam search with keyword biasing* [28], and modifications proposed in the literature, including *cost subtraction* [28], *adaptive boosting* [30], and *Alternate Spelling Prediction (ASP)* [29], in comparison to baseline strategies such as *greedy decoding*, *vanilla beam search*, and *beam search with LM fusion* [10]. Furthermore, we evaluate the effectiveness of biasing methods on three different datasets with varying biasing lists, demonstrating their benefits and drawbacks. We also analyze the methods' effectiveness on *rare* and *Out-of-Vocabulary (OOV)* keyword groups.

The results demonstrate that *LM fusion* can consistently boost the recall and precision of rare and common words, but does not help in recognizing OOV words. We observed significant improvement in recall when using *keyword-biased beam search* along with *cost subtraction* across all three datasets, while the use of the *adaptive boosting* method mainly improved precision and prevented overboosting (boosting a word as a keyword even if it is not actually presented in speech). A further enhancement is achieved by employing an *alternate spelling prediction* approach, which improves keyword recognition of rare and OOV words. Finally, based on our findings, we have provided practical recommendations to make the ASR system more reliable in recognizing special terms and keywords without modifying the acoustic model.

3. **Robust speaker embedding extraction:** Speaker recognition is a broad field of study that addresses two major tasks: **speaker identification** and **speaker verification**. Speaker identification is the task of identifying a person, whereas speaker verification is the task of determining whether the speaker is who they claim to be. In this thesis, we focus on *far-field*, *text-independent* speaker recognition, where the speaker's identity is determined by their speaking style rather than the content of their speech. Typically, such speaker recognition systems operate on unconstrained speech utterances that are converted to a fixed length vector known as *speaker embedding*.

Table 1: Examples of different errors (errors are marked in yellow color).

Error Type	Example
Type a	**Speaker A:** right that's going exactly going back to facebook's optimizer algorithm that's not optimizing for truth right it's optimizing for profit and they they claim to be neutral but of course nothing's neutral right and we have seen the results we've seen what it's actually optimized for and it's not pretty
Type b	**Speaker A:** and presumably you could take all that biased input data and say this high chance recidivism means that we should rehabilitate more i mean like you could take all that same stuff and choose to do a completely different thing with the result of **Speaker B:** the algorithm total that's exactly my point exactly my point you know we could say oh i wonder why people who have this characteristic have so much worse recidivism well let's try to help them find a job maybe that'll help we could use those algorithms those risk scores to try to account for our society
Type c	**Speaker A:** is this a good or bad thing that social media has been able to infiltrate politics

We adopted several approaches from unsupervised domain adaptation to make the speaker recognition models noise-tolerant. In particular, we apply Teacher-Student (TS) [31–33] and Consistency Regularization (CR) [32, 34] techniques on speaker recognition and diarization tasks and compare them with multi-condition training when various noise augmentations are used. The core idea behind CR is to make sure the network produces similar embeddings for the augmented versions of the same unlabeled utterance. It is enforced by an additional regularization term in the loss function:

$$\mathcal{L}_{CR} = \frac{1}{N}\sum_{i=1}^{N} \| f_\theta(\mathcal{A}(x_i)) - f_\theta(\mathcal{A}(x_i)) \|_2^2 \tag{1}$$

where N is a number of examples in the training set, x_i is a ith training example, f_θ is an embedding extractor with parameters θ. By $\mathcal{A}(x)$ we denote a stochastic operation that augments the audio such that speaker identity remains the same. So, the difference is most likely non-zero.

One critical problem with L_{CR} loss is that it is not stable because of unstable target. To mitigate the unstable target problem, the TS model was proposed in [33], where two separate models were used: a Student network with θ parameters and a Teacher with θ' parameters. On unlabeled examples, the Teacher network provides the learning target for the Student network:

$$\mathcal{L}_{TS} = \frac{1}{N}\sum_{i=1}^{N} \| f_\theta^{Student}(\mathcal{A}(x_i)) - f_{\theta'}^{Teacher}(\mathcal{A}(x_i)) \|_2^2 \tag{2}$$

Through the use of teacher-student and consistency regularization, we improved the accuracy of speaker recognition and diarization tasks in noisy and reverberant conditions compared to the baseline model trained with noise augmentations.

4. **Speaker error correction using lexical information:** In practical settings, ASR-SD systems can experience significant degradation in performance due to a variety of factors, including *uniform segmentation with a high temporal resolution*, *inaccurate word timestamps*, *incorrect clustering* and *incorrect estimation of speaker numbers*, as well as *background noise* and *reverberation*. Therefore, it is important to automatically detect these errors and make corrections if possible.

   We conduct an error analysis of an ASR-SD system at the word level, utilizing the Word Diarization Error Rate (WDER) [35] metric to automatically pinpoint words with incorrect speaker labels and

classify these errors into three categories (examples of each type of error are illustrated in Table 1):

(a) Incorrect speaker tags within a paragraph.

(b) The first and last words of a paragraph having incorrect speaker tags.

(c) A complete paragraph being assigned to the wrong speaker.

The leading cause of errors of types (a) and (b) is the use of uniform audio segmentation with a high temporal resolution. Inaccurate ASR word timestamps can also lead to type (b) errors. Type (c) errors typically occur due to inaccurate estimation of the number of speakers and incorrect clustering. Background noise, music and reverberation also contribute to all types of errors.

Furthermore, we proposed two realignment strategies - *language model*-based and *punctuation*-based - to amend errors for words at the boundaries of sentences spoken by different speakers [1]. Both methods improve diarization performance, with the punctuation-based realignment showing the most significant reduction in the word diarization error rate.

5. **Efficient ASR-SD processing for embedded applications:**

   - By reducing the time dimensionality of spectrograms by 4x times, we reduced the computational and memory costs of the attention layers in all following blocks in the neural network, allowing for faster and more efficient ASR processing without compromising accuracy.
   - Quantization in the context of DNN refers to the process of reducing the precision of the weights and activations from floating-point (typically *fp32*) to lower-bit representations, typically integers (typically *int8*). By leveraging dynamic int8 quantization, we successfully decreased the model size 4x times (from 200MB to 50MB) and doubled the inference speed on modern CPUs (equipped with the AVX512 instruction set). This optimization is critical for deployment on resource-constrained devices, offering a practical solution to maintain near-original accuracy while significantly reducing computational demands.
   - Applications such as speech recognition and machine translation use LM to select the most likely translation among many hypotheses. For on-device applications, inference time and model size are just as important as performance. In this work, we explored the fastest family of language models: the N-gram models. We researched the impact of *pruning* (pruning rare N-grams) and *quantization* (using fewer bits to store N-gram counts) methods on model size reduction. Finally, we built a subword language model by using Byte Pair Encoding (BPE) [36].

     We have explored the impact of pruning and quantization on the trade-off between model size and perplexity. Quantization can reduce the size of the model without significantly degrading perplexity. Pruning, on the other hand, drastically reduces the model size at the expense of aggravating perplexity. Moreover, adopting subword tokenization through BPE allows for efficient handling of large vocabularies and unseen words, reducing the model size and improving perplexity scores. Using subword LM, we can prune more aggressively without significant degradation in perplexity compared to word-based LM.

     As a result, we were able to decrease the model size for 10-gram LM from 36.7GB to 100MB.
   - Speaker diarization is based on spectral clustering, which relies on eigen-decomposition. Eigen-decomposition has a high computational complexity. The computational cost of spectral clustering is $O(N^w)$, where $2.37 \leq w \leq 3$ depends on the specific implementation, where N is the number of speaker embeddings. This complexity becomes a significant bottleneck for processing long audio files.

Algorithm 1 Two Stage Clustering Algorithm

Require: Speaker embeddings sequence $x_1, x_2, \cdots, x_N$, Upper limit for spectral clustering U, Number of reduced clusters L ($L \leq U$).
Ensure: Speaker labels $l_1, l_2, \cdots, l_N$

1: **if** $N \leq U$ **then** ▷ Direct spectral clustering if within upper limit
2: $\{l_i\}_{i=1}^{N} \leftarrow$ Apply spectral clustering on $\{x_i\}_{i=1}^{N}$ embeddings
3: **return** $\{l_i\}_{i=1}^{N}$
4: **else** ▷ Pre-cluster to reduce to L clusters
5: Initialize $K = N$
6: **while** $K > L$ **do**
7: Merge two closest clusters
8: $K \leftarrow K - 1$
9: **end while**
10: $\{c_i\}_{i=1}^{L} \leftarrow$ Calculate centroids of L clusters
11: $\{r_i\}_{i=1}^{L} \leftarrow$ Apply spectral clustering on $\{c_i\}_{i=1}^{L}$ centroids
12: **for** $i = 1$ to N **do** ▷ Find the index of the closest centroid to x_i
13: $j \leftarrow \arg\min_k d(x_i, c_k)$ ▷ d is the distance metric, e.g., cosine distance
14: $l_i \leftarrow r_j$
15: **end for**
16: **return** $\{l_i\}_{i=1}^{N}$
17: **end if**

The implementation of a two-stage clustering algorithm for speaker diarization mitigates the computational complexity associated with eigen decomposition. Algorithm 1 describes two-stage clustering algorithm:

- *Steps 1-4:* The algorithm starts by checking if the number of speaker embeddings N is less than or equal to the upper limit U. If this condition is met, it proceeds directly to spectral clustering.
- *Steps 4-9:* If $N > U$, indicating a large number of embeddings, pre-clustering with Agglomerative Clustering is initiated to reduce computational complexity. This process merges the closest clusters iteratively until the number of clusters equals L.
- *Steps 10-11:* After pre-clustering, centroids of the L clusters are calculated. These centroids are then clustered using spectral clustering to determine the cluster labels for the centroids.
- *Steps 12-16:* Finally, each original speaker embedding is assigned to the cluster of the closest centroid, thereby assigning speaker labels to all embeddings.

The AHC pre-clusterer has a computational cost of $O(N^2)$. Thus, the overall computational cost of the clustering step is upper bounded to $O(N^2)+O(L^w)$. By pre-clustering speaker embeddings before applying spectral clustering, we achieved a considerable speed-up, making it more suitable for processing long audio files and enhancing its applicability in real-world scenarios. Specifically, we can achieve a 7x time speed-up for two-hour-long audio.

List of author's publications

1. Karamyan, D., *Adaptive Noise Cancellation For Robust Speech Recognition In Noisy Environments*, Proceedings of the YSU A: Physical and Mathematical Science, XX, xx-yy, (2024).

2. Karamyan, D., Kirakosyan, G., *Keyword-Biased Speech Recognition: A Comparative Study*, Proceedings of NAS RA and NPUA: Series of Technical Sciences, XX, xx-yy, (2023).

3. Karamyan, D., *Text Realignment for Speaker Diarization. Vestnik Of Russian-Armenian University*, Vestnik Of Russian-Armenian University, **1**, 34-43, (2023).

4. Karamyan, D., Kirakosyan, G., *Building a Speaker Diarization System: Lessons from VoxSRC 2023*, Mathematical Problems of Computer Science, **60**, 52–62, (2023).

5. Karamyan, D., Kirakosyan, G., Harutyunyan, S., *Making Speaker Diarization System Noise Tolerant*, Mathematical Problems of Computer Science, **59**, 57–68, (2023).

6. Karamyan, D., Karamyan, T., *Compact N-gram Language Models for Armenian*, Mathematical Problems of Computer Science, **57**, 30–38, (2022).

7. Karamyan, D., Karamyan, T., *A Conformer Based Automated Speech Recognition For Armenian Language*, Scientific Artsakh, 2 (**13**), 224–229, (2022).

References

[1] Paturi, R., Srinivasan, S. & Li, X. Lexical Speaker Error Correction: Leveraging Language Models for Speaker Diarization Error Correction. *Proc. INTERSPEECH 2023*. pp. 3567-3571 (2023)

[2] Wang, Q., Huang, Y., Zhao, G., Clark, E., Xia, W. & Liao, H. Diarizationlm: Speaker diarization post-processing with large language models. *ArXiv Preprint ArXiv:2401.03506*. (2024)

[3] Graves, A., Fernández, S., Gomez, F. & Schmidhuber, J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. *Proceedings Of The 23rd International Conference On Machine Learning*. pp. 369-376 (2006)

[4] Davis, S. & Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions On Acoustics, Speech, And Signal Processing*. **28**, 357-366 (1980)

[5] Abdul, Z. & Al-Talabani, A. Mel Frequency Cepstral Coefficient and its applications: A Review. *IEEE Access*. (2022)

[6] Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y. & Pang, R. Conformer: Convolution-augmented Transformer for Speech Recognition. *Proc. Interspeech 2020*. pp. 5036-5040 (2020)

[7] Rekesh, D., Koluguri, N., Kriman, S., Majumdar, S., Noroozi, V., Huang, H., Hrinchuk, O., Puvvada, K., Kumar, A., Balam, J. & Others Fast conformer with linearly scalable attention for efficient speech recognition. *2023 IEEE Automatic Speech Recognition And Understanding Workshop (ASRU)*. pp. 1-8 (2023)

[8] Radford, A., Kim, J., Xu, T., Brockman, G., McLeavey, C. & Sutskever, I. Robust speech recognition via large-scale weak supervision. *ArXiv Preprint ArXiv:2212.04356*. (2022)

[9] Chan, W., Jaitly, N., Le, Q. & Vinyals, O. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. *2016 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 4960-4964 (2016)

[10] Hannun, A., Maas, A., Jurafsky, D. & Ng, A. First-pass large vocabulary continuous speech recognition using bi-directional recurrent dnns. *ArXiv Preprint ArXiv:1408.2873*. (2014)

[11] Heafield, K. KenLM: Faster and smaller language model queries. *Proceedings Of The Sixth Workshop On Statistical Machine Translation*. pp. 187-197 (2011)

[12] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. & Khudanpur, S. X-vectors: Robust dnn embeddings for speaker recognition. *2018 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 5329-5333 (2018)

[13] Koluguri, N., Li, J., Lavrukhin, V. & Ginsburg, B. SpeakerNet: 1D depth-wise separable convolutional network for text-independent speaker recognition and verification. *ArXiv Preprint ArXiv:2010.12653*. (2020)

[14] Koluguri, N., Park, T. & Ginsburg, B. TitaNet: Neural Model for speaker representation with 1D Depth-wise separable convolutions and global context. *ICASSP 2022-2022 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 8102-8106 (2022)

[15] Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B. & Na, H. ECAPA-TDNN embeddings for speaker diarization. *ArXiv Preprint ArXiv:2104.01466*. (2021)

[16] Jung, J., Heo, H., Yang, I., Shim, H. & Yu, H. A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result. *2018 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 5349-5353 (2018)

[17] Park, T., Kanda, N., Dimitriadis, D., Han, K., Watanabe, S. & Narayanan, S. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*. **72** pp. 101317 (2022)

[18] Von Luxburg, U. A tutorial on spectral clustering. *Statistics And Computing*. **17** pp. 395-416 (2007)

[19] Wang, Q., Downey, C., Wan, L., Mansfield, P. & Moreno, I. Speaker diarization with LSTM. *2018 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 5239-5243 (2018)

[20] Nielsen, F. & Nielsen, F. Hierarchical clustering. *Introduction To HPC With MPI For Data Science*. pp. 195-211 (2016)

[21] Jurafsky, D. & Martin, J. Speech and Language Processing. (2023), https://web.stanford.edu/ jurafsky/slp3/, 3rd edition draft

[22] Karamyan, D. & Kirakosyan, G. Building a Speaker Diarization System: Lessons from VoxSRC 2023. *Mathematical Problems Of Computer Science*. **60** pp. 52-62 (2023)

[23] Wang, Q., Huang, Y., Lu, H., Zhao, G. & Moreno, I. Highly efficient real-time streaming and fully on-device speaker diarization with multi-stage clustering. *ArXiv Preprint ArXiv:2210.13690*. (2022)

[24] Wang, Z., Wang, X., Li, X., Fu, Q. & Yan, Y. Oracle performance investigation of the ideal masks. *2016 IEEE International Workshop On Acoustic Signal Enhancement (IWAENC)*. pp. 1-5 (2016)

[25] Xia, S., Li, H. & Zhang, X. Using Optimal Ratio Mask as Training Target for Supervised Speech Separation. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Kuala Lumpur, Malaysia, Dec. 12–15, 2017. (IEEE,2017)

[26] Lippmann, R., Martin, E. & Paul, D. Multi-style training for robust isolated-word speech recognition. *ICASSP'87. IEEE International Conference On Acoustics, Speech, And Signal Processing*. **12** pp. 705-708 (1987)

[27] Hall, K., Cho, E., Allauzen, C., Beaufays, F., Coccaro, N., Nakajima, K., Riley, M., Roark, B., Rybach, D. & Zhang, L. Composition-based on-the-fly rescoring for salient n-gram biasing. (2015)

[28] Jung, N., Kim, G. & Chung, J. Spell my name: keyword boosted speech recognition. *ICASSP 2022-2022 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 6642-6646 (2022)

[29] Fox, J. & Delworth, N. Improving Contextual Recognition of Rare Words with an Alternate Spelling Prediction Model. *ArXiv Preprint ArXiv:2209.01250*. (2022)

[30] Dingliwal, S., Sunkara, M., Ronanki, S., Farris, J., Kirchhoff, K. & Bodapati, S. Personalization of ctc speech recognition models. *2022 IEEE Spoken Language Technology Workshop (SLT)*. pp. 302-309 (2023)

[31] Mošner, L., Wu, M., Raju, A., Parthasarathi, S., Kumatani, K., Sundaram, S., Maas, R. & Hoffmeister, B. Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning. *ICASSP 2019-2019 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 6475-6479 (2019)

[32] Vanyan, A. & Khachatrian, H. Deep Semi-Supervised Image Classification Algorithms: a Survey.. *J. Univers. Comput. Sci.*. **27**, 1390-1407 (2021)

[33] Tarvainen, A. & Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances In Neural Information Processing Systems*. **30** (2017)

[34] Laine, S. & Aila, T. Temporal ensembling for semi-supervised learning. *ArXiv Preprint ArXiv:1610.02242*. (2016)

[35] Shafey, L., Soltau, H. & Shafran, I. Joint Speech Recognition and Speaker Diarization via Sequence Transduction. *Interspeech*. (2019), https://api.semanticscholar.org/CorpusID:195886036

[36] Sennrich, R., Haddow, B. & Birch, A. Neural machine translation of rare words with subword units. *ArXiv Preprint ArXiv:1508.07909*. (2015)

[37] Karamyan, D. & Kirakosyan, G. The Krisp diarization system for the voxceleb speaker recognition challenge 2023. *The VoxCeleb Speaker Recognition Challenge 2023 (VoxSRC-23)*. (2023)

Ամփոփում

Դավիթ Սպարտակի Քարամյան

Խոսքի հուսալի մշակում՝ ներդրված արհեստական բանականության կիրառություններում

Աշխատանքը նվիրված է խոսքի ճանաչման (speech recognition) և խոսնակների դիարիզացիայի (speaker diarization) համակարգերի կայունության և հուսալիության բարելավման տարբեր մեթոդների մշակմանը, դրանց արդյունավետության գնահատմանը և նշված համակարգերի տեղակայմանը ներդրված կիրառություններում։

Աշխատանքի **հիմնական նպատակներն** են՝

1. Մշակել խոսքի ճանաչման և խոսնակների դիարիզացիայի կայուն և հուսալի համակարգ, մասնավորապես ապահովել այդ համակարգերի կայունությունը՝ տարբեր տեսակի և ուժգնության ֆոնային աղմուկների առկայության դեպքում։

2. Եթե հնարավոր չէ կանխել համակարգի թույլ տված սխալները, ապա առաջարկել և իրականացնել ավտոմատ սխալների ուղղման մեթոդներ։

3. Ուսումնասիրել խոսքի ճանաչման մոդելի ճշգրտությունը առանցքային բառերի դեպքում և առաջարկել մոտեցումներ այդ բառերի ճանաչման ճշգրտությունը բարձրացնելու համար։

4. Գնահատել ստացված համակարգի հաշվողական բարդությունը և մոդելների զբաղեցրած հիշողության ծավալները։ Մշակել օպտիմիզացիոն մեթոդներ՝ համակարգի արագագործությունը բարձրացնելու, իսկ զբաղեցրած հիշողության ծավալը փոքրացնելու համար։ Ուսումնասիրել թե ինչպես են առաջարկված օպտիմիզացիոն մեթոդները ազդում խոսքի ճանաչման և դիարիզացիայի համակարգերի որակական ցուցանիշների վրա։

5. Մշակել առաջարկված մեթոդների համակարգչային ծրագրերը և փորձնական եղանակով համեմատել առաջարկված մեթոդները արդեն գոյություն ունեցող ալգորիթմների հետ, ընդգծել դրանց առավելությունները և թերությունները, ցույց տալ, որ առաջարկված մեթոդները կիրառելի են իրական կյանքի տարբեր տվյալների շտեմարանների վրա։

Ատենախոսության **առաջին գլխում** ներկայացվել են հիմնական օգտագործված մոդելների բնութագրերը և սահմանումները։ Նշվել են առաջադրված խնդիրների արդիականությունը և գործնական կիրառությունները։ Ներկայացվել են ատենախոսական աշխատանքի հիմնական նպատակները և կառուցվածքը։

Երկրորդ գլխում քննարկվել են ձայնի մշակման հիմնական մեթոդները։ Ներկայացվել են ներկայումս առավել հաճախ օգտագործվող խոսքի ճանաչման և խոսնակների դիարիզացիայի համակարգերը։ Քննարկվել են նախկինում կատարված աշխատանքները, որոնք հիմք են հանդիսացել ատենախոսական աշխատանքի համար։

Երրորդ գլուխը նվիրված է խոսքի ճանաչման և խոսնակների դիարիզացիայի համակարգերի կայունության խնդիրներին։ Առաջարկվել է նոր մեթոդ՝ աղմուկի հեռացման (noise cancellation) ալգորիթմի հիման վրա, լավացնելու բառերի ճանաչման ճշգրտությունը աղմուկի պայմաններում։ Ուսումնասիրվել է առանցքային բառերի կանխակալման (keyword biasing) տարբեր մեթոդների արդյունավետությունը, որոնք ուրվագծվել են նախորդ հետազոտություններում և չեն պահանջում որևէ փոփոխություն

խոսքի ճանաչման մոդելում: Աղմուկի և արձագանքի պայմաններում խոսնակների ճանաչման և դիարիզացիայի համակարգերի կայունությունը ապահովելու համար առաջարկվել է վարժեցման երկու ալգորիթմ: Ներկայացվել են իրական տվյալների շտեմարանների վրա իրականացված փորձերի արդյունքները, որոնք ցուցադրում են առաջարկված մեթոդների արդյունավետությունը: Գլխի վերջում ուսումնասիրվել և դասակարգվել են դիարիզացիայի համակարգի թույլ տված հիմնական սխալները և առաջարկվել է երկու ալգորիթմ՝ հիմնված լեզվային մոդելների վրա, ավտոմատ սխալների ուղղման համար:

Չորրոդ գլխում ներկայացված են այն բոլոր մոտեցումները, որոնք կիրառվել են խոսքի ճանաչման և խոսնակների դիարիզացիայի համակարգերի հաշվողական բարդությունը փոքրացնելու համար: Ներկայացված օպտիմիզացիաները հատկապես կարևոր են ներդրված կիրառությունների դեպքում, որտեղ արագագործությունը և զբաղեցրած հիշողության ծավալը կարևոր դեր են խաղում:

Աշխատանքի **գիտական նորույթ** պարունակող առավել կարևոր դրույթները հետևյալն են՝

1. Ներկայացվել են նոր մեթոդներ՝ խոսքի ճանաչման և խոսնակների դիարիզացիայի համակարգերի կայունությունը բարձրացնելու համար՝ տարբեր տեսակի ֆոնային աղմուկների առկայության պայմաններում:

2. Ուսումնասիրվել են առանցքային բառերի կանխակալման տարբեր ալգորիթմների արդյունավետությունը՝ հատուկ անունների և տերմինների ճանաչելու կարողությունը բարելավելու համար:

3. Առաջարկվել է դիարիզացիայի համակարգի թույլ տված սխալների ավտոմատ ուղղման մեթոդներ՝ հիմնված ժամանակակից լեզվային մոդելների վրա:

4. Ուսումնասիրվել և կիրառվել են տարբեր օպտիմիզացման մեթոդներ՝ խոսքի ճանաչման և խոսնակների դիարիզացիայի մոդելների տեղակայումը ներդրված համակարգերում ապահովելու համար:

5. Կատարվել են լայնածավալ փորձարարական աշխատանքներ, որոնք հաստատում են առաջարկված մեթոդների արդյունավետությունը իրական կյանքի տվյալների շտեմարանների վրա: Առաջարկված մեթոդները համեմատվել են բազային այլ մեթոդների հետ, որոնց համեմատ արձանագրվել են զգալի բարելավումներ:

Заключение

Карамян Давид Спартакович

Надежная обработка речи в приложениях встроенного ИИ

Работа посвящена разработке различных методов повышения стабильности и надежности систем распознавания речи (speech recognition) и диаризации дикторов (speaker diarization), оценке их эффективности и внедрению указанных систем в встроенные приложения.

Основными задачами работы являются:

1. Разработать стабильную и надежную систему распознавания речи и диаризации дикторов, в частности обеспечить устойчивость этих систем при наличии фоновых шумов различного типа и интенсивности.

2. Если невозможно предотвратить ошибки, допущенные системой, то предложить и внедрить методы автоматического исправления ошибок.

3. Изучить точность модели распознавания речи в случае ключевых слов и предложить подходы к повышению точности распознавания этих слов.

4. Оценить вычислительную сложность полученной системы и объемы памяти, занимаемые моделями. Разработать методы оптимизации для увеличения скорости работы системы и уменьшения объема занимаемой памяти. Изучить, как предложенные методы оптимизации влияют на показатели качества систем распознавания речи и диаризации.

5. Разработать компьютерные программы по предложенным методам и экспериментально сравнить предлагаемые методы с существующими алгоритмами, выделить их преимущества и недостатки, показать применимость предлагаемых методов к различным реальным базам данных.

Характеристики и определения основных используемых моделей были представлены в **первой главе** диссертации. Отмечена актуальность и практическое применение предложенных задач. Представлены основные цели и структура диссертационной работы.

Во **второй главе** были рассмотрены основные методы обработки звука. Представлены наиболее часто используемые в настоящее время системы распознавания речи и диаризации дикторов. Обсуждались предыдущие работы, послужившие основой диссертационной работы.

Третья глава посвящена проблемам устойчивости систем распознавания речи и диаризации дикторов. Для повышения точности распознавания слов в шумных условиях предложен новый метод, основанный на алгоритме шумоподавления. Изучена эффективность различных методов смещения ключевых слов, которые были изложены в предыдущих исследованиях и не требуют каких-либо изменений в модели распознавания речи. Были предложены два алгоритма обучения, обеспечивающие устойчивость систем распознавания и диаризации дикторов в условиях шума и реверберации. Представлены результаты экспериментов, проведенных на реальных базах данных, которые демонстрируют эффективность предложенных методов. В

конце главы изучены и классифицированы основные ошибки, допускаемые системой диаризации, а также предложены два алгоритма на основе языковых моделей для автоматического исправления ошибок.

В **четвертой главе** представлены все подходы, которые использовались для снижения вычислительной сложности систем распознавания речи и диаризации дикторов. Представленные оптимизации особенно важны для встраиваемых приложений, где важную роль играют скорость и объем памяти.

Научная новизна диссертационной работы заключается в следующем:

1. Представлены новые методы повышения устойчивости систем распознавания речи и диаризации дикторов при наличии различных типов фонового шума.
2. Была изучена эффективность различных алгоритмов смещения ключевых слов для улучшения распознавания собственных имен и терминов.
3. Предложены методы автоматического исправления ошибок, допущенных системой диаризации, на основе современных языковых моделей.
4. Были изучены и применены различные методы оптимизации для внедрения моделей распознавания речи и диаризации дикторов во встроенных системах.
5. Проведены обширные экспериментальные работы, подтверждающие эффективность предложенных методов на реальных базах данных. Предложенные методы сравнивались с другими базовыми методами и были отмечены значительные улучшения.