International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 20 August 2023 **Revised:** 22 September 2023 **Accepted:** 10 October 2023*

Development of efficient techniques for ASR System for Speech Detection and Recognization system using Gaussian Mixture Model- Universal Background Model

[1]Veera V Rama Rao M, [2]Kumar N
[1]Research Scholar, Dept. of CSE, Vels Institute of Science Technology and Advanced Studies, Chennai, India
murali.mvv@gmail.com
[2]Professor, Dept. of CSE, Vels Institute of Science Technology and Advanced Studies, Chennai, India
kumar.se@velsuniv.ac.in

Abstract— Some practical uses of ASR have been implemented, including the transcription of meetings and the usage of smart speakers. It is the process by which speech waves are transformed into text that allows computers to interpret and act upon human speech. Scalable strategies for developing ASR systems in languages where no voice transcriptions or pronunciation dictionaries exist are the primary focus of this work. We first show that the necessity for voice transcription into the target language can be greatly reduced through cross-lingual acoustic model transfer when phonemic pronunciation lexicons exist in the new language. Afterwards, we investigate three approaches to dealing with languages that lack a pronunciation lexicon. Secondly, we have a look at the efficiency of graphemic acoustic model transfer, which makes it easy to build pronunciation dictionaries. Thesis problems can be solved, in part, by investigating optimization strategies for training on huge corpora (such as GA+HMM and DE+HMM). In the training phase of acoustic modelling, the suggested method is applied to traditional methods. Read speech and HMI voice experiments indicated that while each data augmentation strategy alone did not always increase recognition performance, using all three techniques together did. Power normalised cepstral coefficient (PNCC) features are tweaked somewhat in this work to enhance verification accuracy. To increase speaker verification accuracy, we suggest employing multiple "Gaussian Mixture Model-Universal Background Model (GMM-UBM) and SVM classifiers". Importantly, pitch shift data augmentation and multi-task training reduced bias by more than 18% absolute compared to the baseline system for read speech, and applying all three data augmentation techniques during fine tuning reduced bias by more than 7% for HMI speech, while increasing recognition accuracy of both native and non-native Dutch speech.

Keywords- *Gaussian Mixture Model-Universal Background Model, ASR, Power normalised cepstral coefficient***.**

1. Introduction

Because to recent developments in machine learning and speech recognition, millions of individuals around the world now have access to ASR and audio search. Nevertheless, the massive amount of transcribed speech required for constructing high quality ASR engines is not available in most of the 3,000 to 4,000 languages having writing systems [1]. These languages are typically used in less developed or politically and geographically stable regions [2]. Efforts to collect data and provide aid in these areas are prioritised since they are among the most politically volatile and disaster-prone in the globe [3]. As a result of these geopolitical considerations, internet adoption and literacy rates are low, and speech-based communication via cell phones, community radios, or voice messaging has come to dominate. This is because there are not enough audio transcripts available to train effective ASR engines in these languages because so little of the speech is written down [4]. Thus, it is crucial to rapidly roll out ASR capabilities in a wide variety of languages that need minimal or no transcription of spoken language. See Figure 1 for a visual representation of ASR's internal structure.

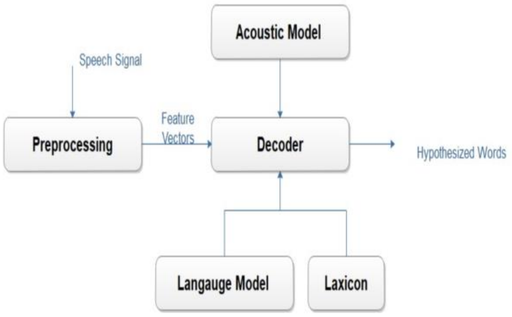


Fig 1: Basic Structure of ASR Module

***Article Received:** 20 August 2023 **Revised:** 22 September 2023 **Accepted:** 10 October 2023*

Pre-processing the input signal to extract features that can be used to model speech is the first step. As part of this procedure, the speech signal is broken down into sub-segments before the features are recovered [5]. Specifically, this research employs Mel-Frequency Cepstral Coefficients (MFCCs) as features, which are derived from the original audio file via windowing, discrete Fourier transform (DFT), logarithm of magnitude, frequency warping on the Mel scale, and inverse discrete consine transform (IDCT) of the log filterbank energies. Finding the most probable word sequences requires an audio model, a language model, and a lexicon that converts written words into phonetic ones [6-9]. For this decoding task, the Viterbi algorithm is frequently utilised in hybrid ASR models.

2. Background

In contrast to the languages, which are typical of those that might be useful in a HADR situation, the speech is not. First, the voiceover sounds like it was recorded professionally in a studio, with clear, legible speech. In certain cases, soft music was played in the background throughout the reading, and artificial echoing was used to make God's voice stand out [10-15]. The majority of the speakers in the corpus are men, but there are some female voices included. There is minimal to no speaker variability in most languages because most recordings are read by only one or two people. Also, the transcripts and audio were automatically split and aligned after discovery of the material. The segment boundaries and transcripts will be inaccurate due to this procedure. Alignment quality was evaluated in [16] by constructing a speech synthesis system from the aligned utterances and then re-synthesizing them. The resulting audio quality was evaluated with the mel-cepstral distortion (MCD) [17], which was a downstream measure of alignment quality. While many languages have good alignment, many others aren't good enough for training purposes. Yet, the subgroup of well-aligned languages in this corpus can provide an attractive testing ground for cross-lingual transmission. The MCD serves as a proxy-metric for alignment quality and is visualised on a map of the world in Figure 2. Each dot represents a language represented in the corpus. In [18], we improved the alignments and phonemic transcripts of 48 languages in the companion corpus by using the zero-resource acoustic modelling approaches discussed in this dissertation [19-22]. This allowed for the first time the systematic investigation of phonetic typology across a wide range of languages.

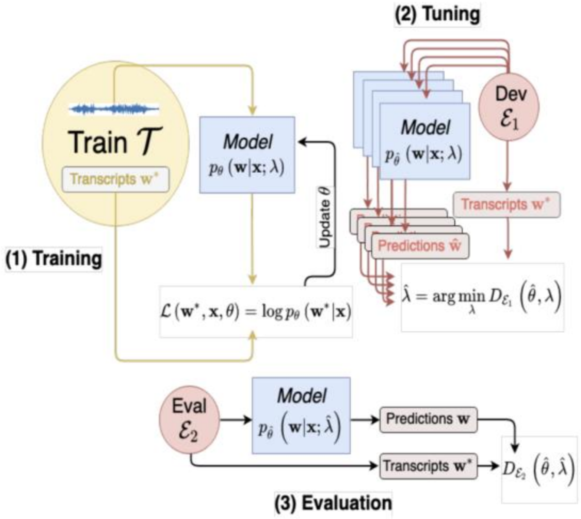

(2) Tuning
Train $\mathcal{T}$
Transcripts $\mathbf{w}^*$
Model
$p_\theta(\mathbf{w}|\mathbf{x};\lambda)$
Update θ
(1) Training
$\mathcal{L}(\mathbf{w}^*,\mathbf{x},\theta) = \log p_\theta(\mathbf{w}^*|\mathbf{x})$
Model
$p_{\hat{\theta}}(\mathbf{w}|\mathbf{x};\lambda)$
Dev $\mathcal{E}_1$
Transcripts $\mathbf{w}^*$
Predictions $\hat{\mathbf{w}}$
$\hat{\lambda} = \arg\min_{\lambda} D_{\mathcal{E}_1}(\hat{\theta},\lambda)$
Eval $\mathcal{E}_2$
Model
$p_{\hat{\theta}}(\mathbf{w}|\mathbf{x};\hat{\lambda})$
Predictions $\mathbf{w}$
Transcripts $\mathbf{w}^*$
$D_{\mathcal{E}_2}(\hat{\theta},\hat{\lambda})$
(3) Evaluation

Fig 2: Process of statistical Analysis of ASR Module

3. Methodology

In numerous practical contexts, automatic speech recognition has shown to be an asset. Despite their widespread use, microphone arrays are not particularly efficient as spatial filters in recordings captured at great distances when there is a lot of reverberation. In many cases, monaural ASR is preferable because of how easy it is to implement. This section examines monoaural ASR's performance under demanding real-world circumstances. CLDNNs are popular monaural audio models [23]. The wide residual BLSTM network (WRBN) in a CLDNN framework performs best for monaural voice recognition in the CHiME-4 challenge's baseline language model. Better LSTM dropout methods and speaker adaption strategies may help RBN. The overfitting issue during LSTM training can be mitigated with the use of dropout for LSTM. An rnnDrop approach is proposed in [24]. for use in speech recognition problems. At each syllable, the dropout mask is sampled and then applied to the cell vector. Dropout masks are sampled in a similar fashion, but they are applied to the input and concealed vectors (Gal dropout).

Figure 3 is a representation of a typical FASR system. The FASR process comprises two stages: the enrolling phase and the testing phase. The enrolling process entails collecting and storing speech samples from the suspect in a database for further comparison. As a matter of terminology, the speech of the suspect is known as training speech and the sample under inquiry is known as test speech. The number of possible liars is high [25]. During the enrolling process, a variety of algorithms are performed to the training speech samples in order to extract a collection of acoustic properties that are unique to each speaker. Extracting the components of a speech sample that are important for speaker recognition and generating characteristics representing unique physical traits for each person constitutes the feature extraction stage. That's why it's so important to get

***Article Received:** 20 August 2023 **Revised:** 22 September 2023 **Accepted:** 10 October 2023*

right when building a speech recognition system. During the modelling or classifier phase, the extracted features are represented in a more compact form by means of appropriate statistical measurements. At this point, the models developed for each possible speaker are saved in a database and later used to evaluate the questioned recording.

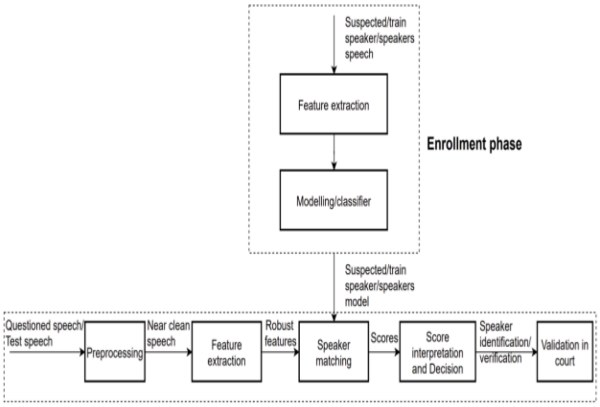

Suspected/train speaker/speakers speech
Feature extraction
Modelling/classifier
Enrollment phase
Suspected/train speaker/speakers model
Questioned speech/ Test speech
Preprocessing
Near clean speech
Feature extraction
Robust features
Speaker matching
Scores
Score interpretation and Decision
Speaker identification/ verification
Validation in court

Fig 3: Flow chart for Speaker Recognization System

Because speaker adaptation yields such large improvements, we propose repeating the adaptation process as shown in Figure 4. Take into account that the RNN language model has been rescored, and that this decoding result represents the final score. Our efforts mirror those of prior work that employed MLLR for iterative adaptation within the framework of LIN training for a DNN-based acoustic model. One can swap the label while leaving all other parameters fixed or stack an additional linear input layer at each iteration to accomplish iterative speaker adaption. The second method assures that the "acoustic model" being adapted is the same as the one provided the adaptation label. This paper compares both strategies.

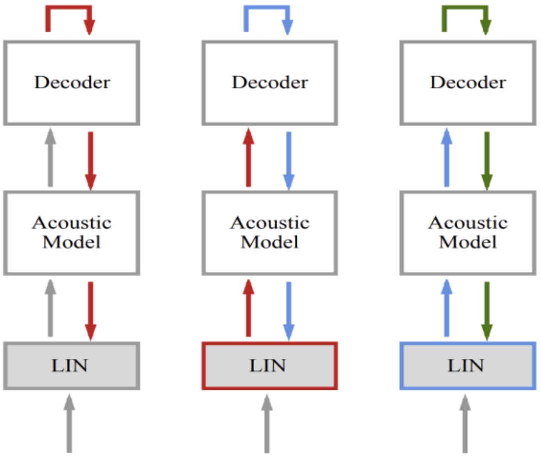


Fig 4: Iterative speaker adaption. Arrow and box colours indicate speaker adaption iterations. Fixes are white while updates are grey.

It is common practise for investigators to initially accuse multiple parties before narrowing the list down to the most likely perpetrator. Due to the numerous distortions that speech experiences, it is highly challenging to choose the correct individual using speech biometric data, which is a one to many speaker verification process. The suggested method involves developing speaker models for the train by fitting the speech data from the train to a UBM. When an SVM classifier is used for further categorization, only the highest-scoring speakers with a grade of "F" are included. While doing multi-class classification, the SVM classifier is typically utilised as a secondary stage. In multiclass classification, it has been noted that although SVM has high discriminative properties, accuracy decreases as more classes are added. Hence, performance can be enhanced by decreasing the number of classes or possible speakers. To achieve this, in the first stage of GMM-UBM classification, we eliminate all but the highest-scoring few speakers from the pool of suspects ('F') for each test speaker. With GMM-UBM, the value of 'F' can be determined from the EER collected during training.

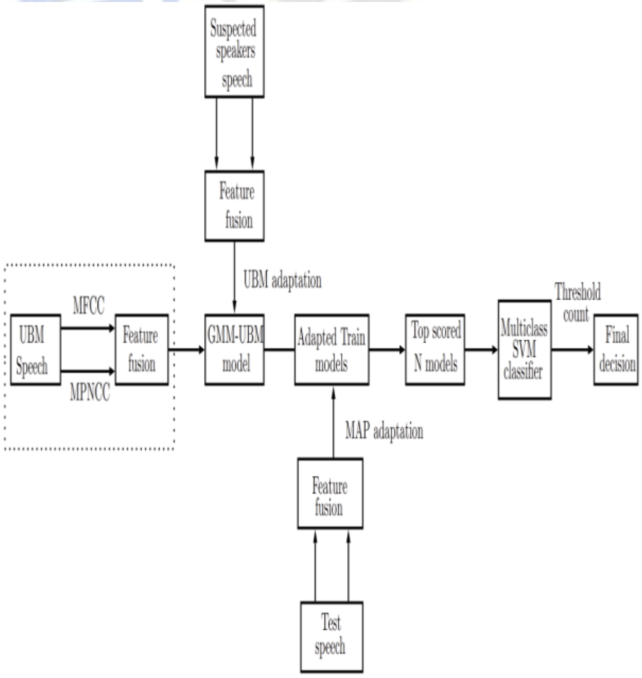


Fig 5: Block diagram for overall proposed system

To verify the speaker, we do not use a score but instead keep track of how many times the test speaker was correctly labelled. There will be FC2 binary classifiers for the 'F' suspected speakers or classes that pass to the next round of classification. Test speech is matched to a specific train speaker if the number of times it has been classified to that speaker is more than a predetermined threshold. When compared to a single-stage classifier, such as the GMM-UBM classifier or the SVM classifier, the error rate is significantly decreased and the number of candidates is reduced. This is due to the fact that the GMM-UBM classifier first excludes a substantial fraction of probable speakers, hence reducing the overall number of error points. Cutting back on the SVM classifier's input class set enhances its overall classification accuracy. Figure 5 is an example of the proposed classifier combination strategy.

Figure 6 depicts a block schematic of the suggested procedure. This paper proposes a novel approach to speaker verification by combining the ideas of feature flipping and affine transformation. For codec-distorted speech, the affine transform approach was proven to increase speaker verification performance. Combining the affine transform technique with the feature-switching idea yields even better results for codec-distorted speech. First, the best feature set is determined for each possible speaker, and a lookup table is established. Using the Euclidean distance between the suspect's clean and affine transformed speech features, the best feature set is chosen.

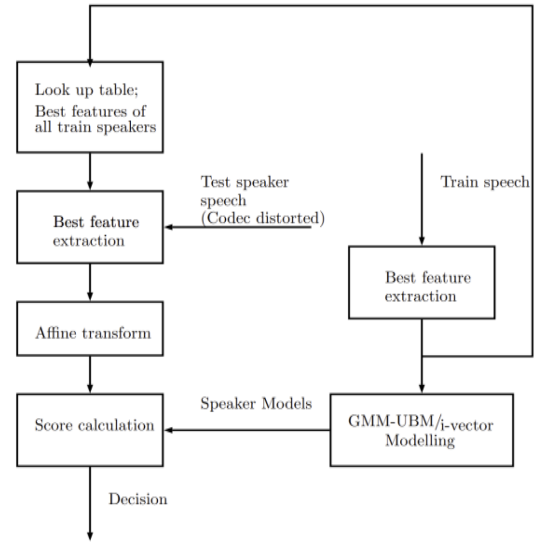


Fig 6: Speaker verification system flow chart

In the Training phase, the best feature of the suspected speaker is taken from the training speech of the suspected speaker, and train models are developed by GMM-UBM/I vector modelling.

After obtaining a test speech in the testing phase, it must be compared to the recorded voices of all possible culprits. Then, the lookup table is used to determine which aspect of the test speaker's voice is most indicative of the suspect speaker. In order to compensate for the fact that the test speaker's voice has been warped by the codec, the characteristics extracted from the speaker's speech are affinely transformed. To score and validate, we employ the suspect speaker model with the most desirable characteristics. Each suspect speaker must go through the aforementioned verification process a second time after the best element of their speech has been extracted from the test speaker's speech. If the best feature of the suspect speaker can be determined through testing, then applying an affine transformation to that feature will result in higher speaker verification ratings than would be achieved otherwise.

4. Results

In Table 1, we can see how SSUSI and PIT stack up against one another in terms of SDR and WER. SSUSI outperforms PIT in both SDR and WER because of its capacity to make use of speaker information. With only 32 candidate profiles in the speaker inventory, SSUSI still achieves an SDR of 10.8 dB, which is significantly higher than PIT's SDR of 8.7 dB in the situation of 30 irrelevant profiles. Keep in mind that SSUSI only uses 2 completely unrelated profiles for training. It is clear that SSUSI is reliable due to the outcomes with 6, 22, and 30 irrelevant profiles. On the assumption of an irrelevant profile of zero, SSUSI achieves a 48% relative improvement in WERs over PIT. There is still a relative increase of 34% even though there are 30 unnecessary profiles. Table 1 shows that the WERs for the LibriSpeech corpus are all above average. It's because of the inaccuracies in estimated speech that cause this.

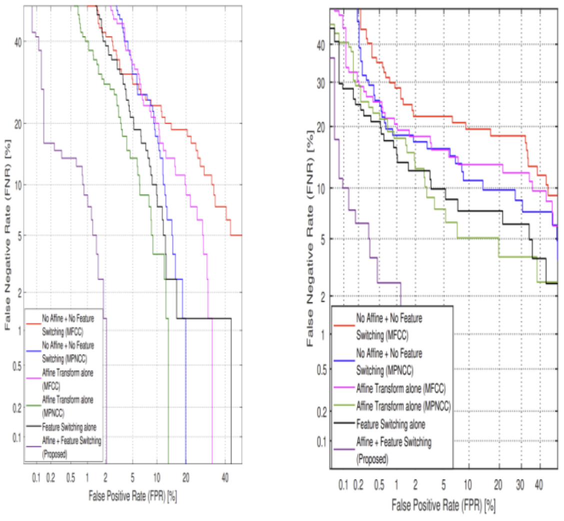
Table 1: Table for Speech Extraction and SSUSI of different parameters

Parameter	Speech Extraction			SSUSI		
ir-profiles	0	1	2	0	1	2
SDR (dB)	11.5	11.1	10.9	12.2	12	11.9
WER (%)	21.9	23.3	24.4	19.1	19.9	20.4

Distances in Euclidean space between 10 speakers' initial MFCC and MPNCC features and their affine transformations.

Table 2: MFCC and MPNCC features and their affine transformations.

Speaker	MPNCC Features	MFCC Features
Speaker 10	2.58	2.38
Speaker 9	3.13	4.78
Speaker 8	4.43	3.62
Speaker 7	3.61	4.89
Speaker 6	2.76	1.70
Speaker 5	0.68	1.04
Speaker 4	2.06	1.98
Speaker 3	2.50	5.19
Speaker 2	2.60	3.43
Speaker 1	4.01	3.64



(a) (b)

Fig 7: Plot for Speaker verification system (a) i-Vector System (b) GMM-UBM system

In Table 3, a comparison is made between the suggested method for speaker verification and the GMM-UBM and i-vector systems that are contained within the TIMIT database. In conclusion, the simulation results from the GMM-UBM and i-vector systems are compared to the results from the most advanced x-vector system. This comparison demonstrates that the proposed method is superior to the state-of-the-art system. The comparative results for the TIMIT and VoxCeleb1 databases are shown in Figure 7a and 7b, respectively, and a summary of the findings can be found in Table 3. The basic x-vector system is outperformed by the suggested solution for speaker verification when using codec-distorted speech in both the GMM-UBM and i-vector systems.

Table 3: comparison table for GMM-UBM and i-vector system

Methods	GMM-UBM System(EER(%))	i-vector system (EER(%))
Without affine transform feature switching (MPNCC)	12.78	11.45
MFCC Features "Without affine transform feature switching"	19.57	16.89
MFCC Features "With affine transform feature switching"	13.57	12.73
With feature transform alone	9.49	7.24
MPNCC "With affine transform feature switching"	8.50	5.29
Proposed Method "With affine transform and feature switching"	1.91	1.2

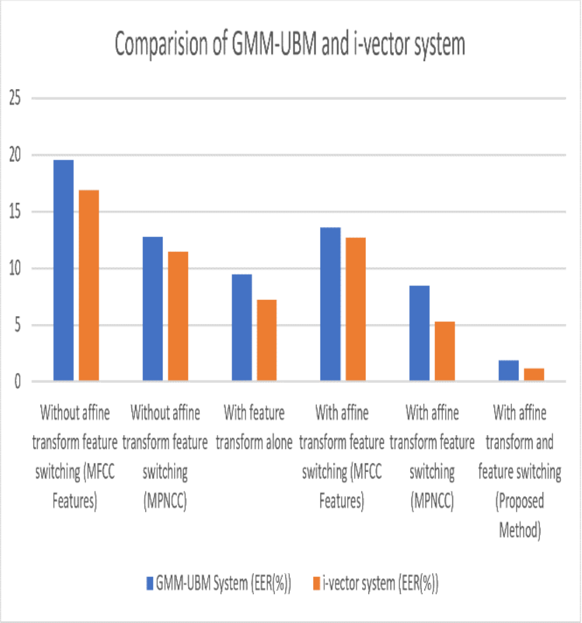


Fig 8: Plot for comparison of GMM-UBM and i-vector system

5. Conclusion

The usage of ASR in practical settings is widespread, and examples include smart speakers and meeting transcription. Most current ASR systems use cloud-based servers to process audio signals. Because of this, leaks and attacks on personally identifiable information included in speech are possible. In this chapter, we offer a unique method for speaker verification using codec-distorted audio by combining feature flipping with affine transform. It is taken into account that the CELP codec will distort any speech retrieved through a mobile communication, and so speaker verification from such distorted audio must be performed. Using the TIMIT database, the proposed method achieved an EER of 1.85% when using the GMM-UBM classifier and 1.16% when using the i-vector classifier. These percentages are 7.23 and 5.68 for the VoxCeleb1 database. Compared to both state-of-the-art x-vector systems and other competing approaches, these error rates are significantly lower. As a result, the suggested approach works wonderfully as a forensic speaker verification tool. However, model retraining necessitates a substantial amount of computational resources and the availability of training data, limiting the usefulness of model compression strategies. In this research, we look into the possibility of using model compression for ASR models without resorting to model retraining. Recently, the DNN method has been investigated in conjunction with the HMM classifier technique; however, the inclusion of model space adaption strategies, such as the CNN and RNN methods, would greatly improve the quality of the work. These methods have attracted a lot of interest due to their ability to replicate the DNN task.

References

[1] M. Yousefi and J. H. L. Hansen, "Block-Based High Performance CNN Architectures for Frame-Level Overlapping Speech Detection," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 28-40, 2021, doi: 10.1109/TASLP.2020.3036237.

[2] I. Hwang and J. -H. Chang, "End-to-End Speech Endpoint Detection Utilizing Acoustic and Language Modeling Knowledge for Online Low-Latency Speech Recognition," in IEEE Access, vol. 8, pp. 161109-161123, 2020, doi: 10.1109/ACCESS.2020.3020696.

[3] F. Tao and C. Busso, "End-to-End Audiovisual Speech Recognition System With Multitask Learning," in IEEE Transactions on Multimedia, vol. 23, pp. 1-11, 2021, doi: 10.1109/TMM.2020.2975922.

[4] H. Dinkel, S. Wang, X. Xu, M. Wu and K. Yu, "Voice Activity Detection in the Wild: A Data-Driven Approach Using Teacher-Student Training," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 1542-1555, 2021, doi: 10.1109/TASLP.2021.3073596.

[5] S. Latif, J. Qadir, A. Qayyum, M. Usama and S. Younis, "Speech Technology for Healthcare: Opportunities, Challenges, and State of the Art," in IEEE Reviews in Biomedical Engineering, vol. 14, pp. 342-356, 2021, doi: 10.1109/RBME.2020.3006860.

[6] A. S. B. Wazir, H. A. Karim, H. S. Lyn, M. F. Ahmad Fauzi, S. Mansor and M. H. Lye, "Deep Learning-Based Detection of Inappropriate Speech Content for Film Censorship," in IEEE Access, vol. 10, pp. 101697-101715, 2022, doi: 10.1109/ACCESS.2022.3208921.

[7] J. -W. Kim, H. Yoon and H. -Y. Jung, "Linguistic-Coupled Age-to-Age Voice Translation to Improve Speech Recognition Performance in Real Environments," in IEEE Access, vol. 9, pp.136476-136486,2021,doi: 10.1109/ACCESS.2021.3115608.

[8] E. Egorova, H. K. Vydana, L. Burget and J. H. Černocký, "Spelling-Aware Word-Based End-to-End ASR," in IEEE Signal Processing Letters, vol. 29, pp. 1729-1733, 2022, doi: 10.1109/LSP.2022.3192199.

[9] M. Price, J. Glass and A. P. Chandrakasan, "A Low-Power Speech Recognizer and Voice Activity Detector Using Deep Neural Networks," in IEEE Journal of Solid-State Circuits, vol. 53, no. 1, pp. 66-75, Jan. 2018, doi: 10.1109/JSSC.2017.2752838.

[10] L. Sari, M. Hasegawa-Johnson and S. Thomas, "Auxiliary Networks for Joint Speaker Adaptation and Speaker Change Detection," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 324-333, 2021, doi: 10.1109/TASLP.2020.3040626.

[11] T. Tambe et al., "A 16-nm SoC for Noise-Robust Speech and NLP Edge AI Inference With Bayesian Sound Source Separation and Attention-Based DNNs," in IEEE Journal of Solid-State Circuits, vol. 58, no. 2, pp. 569-581, Feb. 2023, doi: 10.1109/JSSC.2022.3179303.

[12] Q. Liu, Z. Chen, H. Li, M. Huang, Y. Lu and K. Yu, "Modular End-to-End Automatic Speech Recognition Framework for Acoustic-to-Word Model," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2174-2183, 2020, doi: 10.1109/TASLP.2020.3009477.

[13] Y. Lin, L. Deng, Z. Chen, X. Wu, J. Zhang and B. Yang, "A Real-Time ATC Safety Monitoring Framework Using a Deep Learning Approach," in IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 11, pp. 4572-4581, Nov. 2020, doi: 10.1109/TITS.2019.2940992.

[14] B. Yusuf, B. Gundogdu and M. Saraclar, "Low Resource Keyword Search With Synthesized Crosslingual Exemplars," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 7, pp. 1126-1135, July 2019, doi: 10.1109/TASLP.2019.2911164.

[15] X. Du and C. -M. Pun, "Robust Audio Patch Attacks Using Physical Sample Simulation and Adversarial Patch Noise Generation," in IEEE Transactions on Multimedia, vol. 24, pp. 4381-4393, 2022, doi: 10.1109/TMM.2021.3116426.

[16] S. Lu, J. Lu, J. Lin and Z. Wang, "A Hardware-Oriented and Memory-Efficient Method for CTC Decoding," in IEEE Access, vol. 7, pp. 120681-120694, 2019, doi: 10.1109/ACCESS.2019.2937680.

[17] C. Wang et al., "ARoBERT: An ASR Robust Pre-Trained Language Model for Spoken Language Understanding," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1207-1218, 2022, doi: 10.1109/TASLP.2022.3153268.

[18] Y. Liu, T. Lee, T. Law and K. Y. -S. Lee, "Acoustical Assessment of Voice Disorder With Continuous Speech Using ASR Posterior Features," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 6, pp. 1047-1059, June 2019, doi: 10.1109/TASLP.2019.2905778.

[19] K. Deng, G. Cheng, R. Yang and Y. Yan, "Alleviating ASR Long-Tailed Problem by Decoupling the Learning of Representation and Classification," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 340-354, 2022, doi: 10.1109/TASLP.2021.3138707.

[20] L. -D. Van, Y. -C. Tu, C. -Y. Chang, H. -J. Wang and T. -P. Jung, "Hardware-Oriented Memory-Limited Online Artifact Subspace Reconstruction (HMO-ASR) Algorithm," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 68, no. 12, pp. 3493-3497, Dec. 2021, doi: 10.1109/TCSII.2021.3124253.

[21] G. Cheng, H. Miao, R. Yang, K. Deng and Y. Yan, "ETEH: Unified Attention-Based End-to-End ASR and KWS

Architecture," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1360-1373, 2022, doi: 10.1109/TASLP.2022.3161159.

[22]R. Yang, G. Cheng, H. Miao, T. Li, P. Zhang and Y. Yan, "Keyword Search Using Attention-Based End-to-End ASR and Frame-Synchronous Phoneme Alignments," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3202-3215, 2021, doi: 10.1109/TASLP.2021.3120632.

[23] J. Tang, J. Zhang, Y. Song, I. McLoughlin and L. -R. Dai, "Multi-Granularity Sequence Alignment Mapping for Encoder-Decoder Based End-to-End ASR," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 2816-2828, 2021, doi: 10.1109/TASLP.2021.3101921.

[24] G. Sun, C. Zhang and P. C. Woodland, "Minimising Biasing Word Errors for Contextual ASR With the Tree-Constrained Pointer Generator," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 345-354, 2023, doi: 10.1109/TASLP.2022.3224286.

[25] Y. Higuchi, N. Moritz, J. Le Roux and T. Hori, "Momentum Pseudo-Labeling: Semi-Supervised ASR With Continuously Improving Pseudo-Labels," in IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1424-1438, Oct. 2022, doi: 10.1109/JSTSP.2022.3195367.