Learning linguistic content embeddings for phrase and language independent utterance verification

Mohammad Mohammadamini, Nathan Griot, Driss Matrouf

▶ **To cite this version:**

Mohammad Mohammadamini, Nathan Griot, Driss Matrouf. Learning linguistic content embeddings for phrase and language independent utterance verification. 2024. hal-04478250

HAL Id: hal-04478250
https://hal.science/hal-04478250

Preprint submitted on 26 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Learning linguistic content embeddings for phrase and language independent utterance verification

Mohammad Mohammadamini, Nathan Griot, Driss Matrouf

LIA (Laboratoire Informatique d'Avignon)
University of Avignon, France
mohammad.mohammadamini, nathan.griot, driss.matrouf@univ-avignon.fr

Abstract

In this paper, we introduce a phrase and language-independent utterance verification system. The objective of an utterance verification system is to confirm whether the linguistic content of two speech utterances is the same or not. Our proposed approach is phrase and language-independent, therefore it can generalize for unseen phrases and even unseen languages. The proposed framework is based on a ResNet embedding extractor trained on the Common Voice dataset which is optimized by a linguistic content classification task.The proposed approach can be used in text-dependent speaker recognition systems, passphrase verification systems, and keyword spotting systems. Our system is tested on several protocols including Deepmine for text-dependent speaker verification and Speech Commands' keyword-spotting benchmarks. Obtaining high performance for unseen phrases and languages makes our approach plausible for utterance verification in low-resource and even zero-resource languages. For example, the EER for a protocol on Common Voice English is 0.16 while for Common Voices French without using French training data the EER is 1.23.

Key terms: utterance verification, linguistic-content embeddings, text-dependent speaker verification, keyword spotting

1. Introduction

A phrase recognizer or an utterance verification system confirms whether two utterances contain the same linguistic content or not [1, 2]. Utterance verification systems can be utilized in various speech technologies such as text-dependent speaker recognition systems, passphrase verification systems, command systems, and keyword spotting systems. A speech signal encodes different kinds of information such as speaker, language, emotion, and linguistic content. Obtaining disentangled representations that encode a particular aspect of information is used in different speech technologies such as speaker recognition [3], language identification [4], and emotion recognition [5]. In this paper, we propose a new approach for utterance verification that is based on fixed-length embeddings that capture linguistic content.

The previous work on phrase/utterance verification is mainly limited to a close set of sentences [1, 2]. In a close-set setup, the verification of new utterances requires preparing new data and system modification which leads to less flexibility, less reusability, and more time and money consumption. In this paper, we introduce a phrase and language-independent utterance verification system that extends the previous work in several ways. The proposed system makes it possible to generate an embedding characterizing the linguistic content regardless of the sentence pronounced, even when it has never been seen in the training data. From this perspective, we call the proposed system *phrase-independent*. Furthermore, we will show that with a relatively small performance degradation, it can be used even for new languages without training data for that language. In this sense, we call our approach *language-independent*.

A common application for an utterance verification system is speaker recognition systems. Utterance verification is used in text-dependent speaker recognition systems in two ways. Firstly it can be used to filter out the trials containing different linguistic content. By doing so, the linguistic variability can be controlled which leads to higher performance in severe acoustical conditions [6, 7]. Also, passphrase verification can be used as a secondary metric alongside the speaker characteristics to make the speaker recognition systems more robust [8]. The previous work in this direction is mainly focused on passphrase verification for a small pool of sentences. The RSR2015 [6] contains 30 unique sentences in the training part; Deepmine is another well-known dataset that contains 5 English and 5 Persian phrases for text-dependent speaker recognition [9]. Our proposed approach can exempt us from the time and money-consuming tasks of preparing text-dependent training datasets.

Keyword spotting systems and command systems are becoming more popular in speech assistants. The goal of such systems is to recognize specific content in an audio stream to process a command [10]. The main configuration of keyword spotting systems is based on classifying a close set of words/phrases [10]. Similar to text-dependent speaker recognition systems the benchmarks are limited to a small number of sentences or words [11]. For example, *Speech Commands* is a known keyword spotting benchmark that contains 35 unique words commonly used as commands in speech assistants. Also, the majority of datasets are devoted to high-resource languages such as English or Mandarin [10]. The fact that our approach is phrase-independent and language-independent makes it a more flexible and reusable keyword-spotting system. Indeed, the users can define their commands for voice assistants without having the written or transcribed data in their language. Also, it lets them personalize the commands for the voice assistants.

Wake word detection is a particular command recognition system that aims to activate a speech assistant by a specific command like "Hi Alexa" [12, 13]. In practice, such systems run on low-resource devices and listen continuously for a specific wake word. Similar to the aforementioned applications wake-word detection is constrained to recognizing a close set of phrases which makes the use of speech assistance more difficult for non-English speakers or underrepresented languages.

As it is described, the majority of available utterance verification systems are constrained to a small close set of phrases.

In this paper, we propose a general phrase and language-independent utterance verification system by training discriminant linguistic content embeddings. In a similar approach to text-independent speaker recognition systems, we train a ResNet embedding extractor that captures discriminant linguistic content representations at the utterance level. To do so, it is required to have a transcribed/labeled speech dataset with multiple repetitions for each utterance. Driving a dataset for utterance verification from Common Voice is another contribution of this work. The details about the training dataset are presented in Section 3. To the best of our knowledge, this methodology is not previously used to achieve linguistic content representations for speech utterances.

The rest of the paper is organized as: in section 2 the configuration of the proposed utterance verification system and the architecture of the embedding extractor are described. In section 3 the training and evaluation datasets are presented, and section 4 discusses the obtained results.

2. Utterance verification system

Different kinds of information such as speaker characteristics, recording devices, recording environment, and linguistic content are encoded in a speech signal. Finding a compact representation for a variable-length speech signal for different tasks such as speaker recognition [3], language identification [4], and emotion recognition [5] is explored. In this paper, we propose an approach to learn representations that captures linguistically discriminant characteristics. The idea of the proposed system is taken from speaker embedding extractions [3, 14, 15]. Our proposed system is depicted in Figure 1. The training and application parts of the proposed configuration are described in the following.

2.1. System configuration

2.1.1. Training phase

The training phase is composed of three steps. Firstly the acoustic features are extracted. After that, we have an utterance embedding extractor. In our case, we used ResNet34 which is described in Table 1. Finally, there is an utterance classification that classifies the utterances based on their linguistic content. By imposing on the output of the embedding extractor to generate linguistically discriminant representations, the embedding extractor captures the linguistic content information.

2.1.2. Application phase

In the application phase, we use the trained utterance embedding extractor to generate a linguistic content discriminant representation. The obtained representations are expected to encode the linguistic content characteristics independent of their language and spoken phrase. In this step, the classifier is removed and the given representations by the trained embedding extractor are used for utterance verification. For a pair of given embeddings, a decision on their sameness in terms of content will be taken based on the similarity between the embeddings. We are using the cosine distance to compare each pair of representations, if this distance is smaller than a threshold the embeddings are considered to have the same linguistic content, otherwise, they will be considered to have different content.

As we mentioned in Section 1, the proposed utterance verification can be used for several applications including passphrase verification, text-dependent speaker recognition, command recognition, and word wake detection.

- **Passphrase verification:** Authenticating the user's linguistic content as a secondary metric alongside the speaker characteristics can improve the reliability of speaker recognition systems. In this application for each registered user, both speaker characteristics and spoken phrases (passphrases) should be authenticated. Since our proposed approach is phrase-independent, the users can change their passphrase. In this configuration, if the passphrase is not revealed the speaker recognition system remains safe facing some spoofing attacks including the replaying attack, voice conversion, and personification attacks [16].
- **Text-dependent speaker verification:** In text-dependent speaker recognition systems, only specific phrases are accepted that lead to reducing the linguistic variability and higher performance of speaker recognition systems [6]. Our utterance verification systems can be used as a preprocessing step to filter out the undefined phrases in the system.
- **Command recognition:** In this case, there is a set of registered commands [11]. Our proposed approach can support both open and closed set commands. Therefore, the user can add or customize the voice commands. The generalizability of the proposed approach to new languages without having training data can make the speech assistants more user-friendly and available for a bigger community.
- **Word wake detection:** In this case there is normally one registered command that wakes up the speech assistant. This configuration is a special case of a command recognition system. Having a phrase-independent utterance verification system, the user can define a personalized wake word for the speech assistants.

2.2. Speaker embedding extractor

The embedding extractor used in this paper is a variant based on ResNet [14]. The ResNet model for extracting embeddings consists of three modules: a set of ResNet Blocks, a statistics-level layer, and segment-level representation layers.

- ResNet (*Residual Network)* uses stacks of many Residual Blocks. A Residual Block is made up of two 2- two-dimensional convolutional Neural Network (CNN) layers separated by a non-linearity (ReLU). The input of the residual block is added to its output to constitute the input of the next block.
- The *statistics-level* component is an essential component in converting a variable-length speech signal into a single fixed-dimensional vector. We are using the attentive statistics pooling [17], which aggregates over frame-level output vectors of the DNN and computes a weighted mean and weighted standard deviation.
- The *segment-level* component maps the segment-level vector to the utterance class. The weighted mean and weighted standard deviation are concatenated together and forwarded to the next layers and finally to the softmax output layer.

The detailed topology of the used ResNet is shown in Table 1. Batch-norm and ReLU layers are not shown. The dimensions are (Frequency×Channels×Time). The input is comprised of

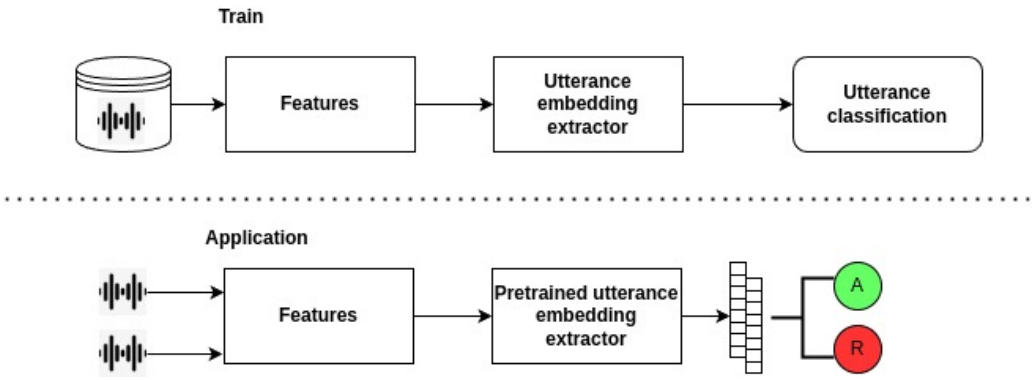

Figure 1: *The configuration of proposed Phrase and language-independent utterance verification system*

Table 1: *The baseline ResNet-34 architecture.The last row, Y is the number of unique utterances. The dimensions are (Frequency×Channels×Time). The input is comprised of 60 filter banks.*

Layer name	Structure	Output
Input	–	$60 \times 400 \times 1$
Conv2D-1	3×3, Stride 1	$60 \times 400 \times 32$
ResNetBlock-1	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 3$, Stride 1	$60 \times 400 \times 32$
ResNetBlock-2	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 4$, Stride 2	$30 \times 200 \times 64$
ResNetBlock-3	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 6$, Stride 2	$15 \times 100 \times 128$
ResNetBlock-4	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 3$, Stride 2	$8 \times 50 \times 256$
Flatten	–	8×2048
Pooling	–	2048
Dense1	–	256
Dense2 (Softmax)	–	Y
Total	–	–

60 filter banks from speech segments. The ResNet is trained with the Additive Angular Margin Loss (ArCFace)[18] function (Equation. 1). For n number of training examples and y number of unique utterances, the network is trained to minimize:

$$\mathcal{L}_{ArcFace} = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{e^{s}\cdot(\cos\theta_{y_i}+m)}{e^{s.\cos}(\theta_{y_i}+m)+\sum_{j\neq y_i}e^{s.(\cos\theta_j)}} \quad (1)$$

Where y_i is the ith utterance, s is a scale factor and m is the margin.

3. Dataset

3.1. Training data

The majority of utterance-verification research is done on limited task-specific datasets such as text-dependent speaker recognition [2, 19, 20, 21]. The RSR2015 [6] and Deepmine [9] are two common datasets used in the literature. In this section, we introduce a new dataset driven from Common Voice 15 to train a general-purpose phrase and language-independent utterance verification system.

Training the proposed linguistic content embedding extractor needs a big transcribed/labeled speech dataset with several repetitions per sentence. Common Voice is among the rare speech recognition datasets that have this specification partially. In our research, we drive the training dataset from the English part of Common Voice 15 [1]. To achieve a relatively balanced training dataset we chose sentences that have more than 5 and less than 100 repetitions. The resulting dataset comprises 15,455 unique sentences. The total number of utterances is about $505k$ and 416 hours of speech data. The training data specifications are presented in table 2.

Table 2: *The Common Voice training data specifications for utterances with more than 5 repetitions per sentence.*

Item	N
Unique utterances	15455
Utterance repetitions	>5
Total Utterances	$\approx 505k$
Total Duration	416h

3.2. Evaluation data

We used several evaluation datasets that are described in this subsection. In evaluation protocols, *target* trials are pairs of utterances with the same linguistic content, and *nontarget* trials are pairs with different linguistic content.

- **Common Voice English:** In this protocol, 1740 unique utterances with 4 or 5 repetitions for each phrase from the English part of Common Voice 15 are chosen. There is no overlap between the chosen files and the training part described in section 3 in terms of linguistic content. From these files, 773k trials are generated.
- **Common Voice French:** In this protocol 225 unique files with 10 repetitions from the French part of Common Voice 15 are chosen. In this protocol, there are 381k trials. The goal of this protocol is to show the generalization of the proposed system for new languages.
- **Deepmine English:** Deepmine is a dataset for text-dependent speaker recognition. In this protocol, there are 27k trials generated from the recording of 5 unique phrases in the Deepmine dataset. This protocol is a case study of our approach to text-dependent speaker recognition [9].

[1] https://commonvoice.mozilla.org/en/datasets

- **Deepmine Persian:** In this protocol there are 25.4k trials generated from the Persian part of the Deepmine corpus which includes different recordings for 5 unique utterances [9]. This protocol shows the generalizability to another new language.
- **Speech Commands:** *Speech Commands* is a widely used dataset for keyword spotting recognition. In this protocol, each utterance contains a single word, and the total number of unique utterances is 35. This protocol is created from 2000 randomly chosen files from the evaluation part of the Speech Commands dataset [11].

Table 3 summarizes the details of evaluation protocols. The last columns show the average duration of test files in seconds.

Table 3: *Evaluation protocols specifications.*

Protocol	Task	trials	utterances	duration(s)
Common Voice EN	General	773k	1740	3.25
Common Voice Fr	General	381k	225	3.47
Deepmine En	TDSV	27k	5	2.30
Deepmine Fa	TDSV	25.4k	5	1.9
Speech command	KWS	1m	35	0.73

4. Results and Discussion

The utterance embedding extractor is trained with driven data from Common Voice 15 discussed in section 3. For each clean file, four augmented versions are created which leads to $2.5m$ samples in the training data. The training data is augmented with different branches of Musan [22] and RIR files [3] using the Kaldi toolkit [23]. The embedding extractor is optimized in $2,000$ iterations, with a learning rate started by 0.2 with a decay rate of 10^{-4}. The size of each mini-batch is 128 and the mini-batch files are chosen randomly.

In the test step, the Equal Error Rate (EER) is calculated based on the cosine distance between pairs of files. If the cosine distance is lower than a threshold the files will be considered as same otherwise they have different linguistic content.

4.0.1. Common Voice protocols

As it is presented in Table 4, the EER is 0.16 on the Common Voice English protocol. The results show a high performance of the proposed approach. In the second experiment on Common Voice French, it is shown that extracted embeddings capture discriminant language characteristics regardless of the language. In this experiment, the EER is 1.2. However, there is relative performance degradation, but even without fine-tuning, it is a plausible result.

4.0.2. Utterance verification for text-dependent speaker recognition

In the second group of experiments; we did utterance verification on the Deepmine corpus which is a text-dependent speaker recognition. Similar to Common Voice protocols, the EER for the English version is 0.16, and for the Persian part, it is 1.9 (Table 4). Since, the Persian language and English versions are very different in terms of lexicon, syntax, and morphology, the obtained results prove the feasibility of repeating our approach for other new languages.

Table 4: *Obtained results for evaluation protocols in EER.*

Protocol	EER
Common Voice En	0.16
Common Voice Fr	1.23
Deepmine En	0.13
Deepmine Fa	1.9
Speech command	6.00

4.0.3. Utterance verification for command verification

Our last experiment is devoted to a command verification protocol created on *Speech Commands*. In this experiment, the EER increased is 6.0. The obtained result on this dataset is competitive in comparison to reported results on the same dataset [11]. To have a more precise interpretation of the errors in this protocol we observed that the majority of errors come from trials where the only phonological difference between two files is one consonant. For example, there are pairs of "three" and "tree" or "on", and "off" trials which are the main resource of the error.

4.0.4. Robustness to duration

In other domains such as speaker recognition, the weakness of embeddings for short-duration utterances is significant [24]. We see that in linguistic content embeddings, the duration doesn't impact the results significantly. In another experiment, the Common Voice En protocol is evaluated for test files with different durations. As shown in Table 5, the EER of short duration between $[1, 2]$ seconds is almost the same as files in the range of $[4, 5]$ seconds.

Table 5: *The impact of duration variability*

Duration(s)	EER
$[1, 2]$	0.15
$[2, 3]$	0.13
$[3, 4]$	0.12
$[4, 5]$	0.16

5. Conclusion

In this paper, we proposed an utterance verification system based on discriminant linguistic content embeddings. The proposed system is phrase-independent and it can generalize to a high degree to new languages. Our approach is performing well in different tasks such as text-dependent speaker recognition and keyword spotting applications. This work can be extended in several ways. The same idea can be tested in the case of having several languages in the training data to capture a broader phonetic variability. Besides the English version of Common Voice 15, a part of Welsh, German, Persian, and Kabyle languages have several repetitions per sentence that can help to foster the research in this direction. Also replacing the font-end features with a self-supervised representation such as Wav2Vec [25] or Whisper [26] is expected to capture linguistic characteristics more explicitly.

6. References

[1] Gilles Boulianne, "Language-independent voice passphrase verification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4490–4494.

[2] Hossein Zeinali, Lukas Burget, Hossein Sameti, and Honza Cernocky, "Spoken Pass-Phrase Verification in the i-vector Space ," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 372–377.

[3] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[4] David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "Spoken Language Recognition using X-vectors," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 105–111.

[5] Raghavendra Pappagari, Tianzi Wang, Jesus Villalba, Nanxin Chen, and Najim Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7169–7173.

[6] Anthony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "Text-dependent speaker verification: Classifiers, databases and rsr2015," *Speech Communication*, vol. 60, pp. 56–77, 2014.

[7] Tomi Kinnunen, Md. Sahidullah, Ivan Kukanov, Héctor Delgado, Massimiliano Todisco, Achintya Kr. Sarkar, Nicolai Bæk Thomsen, Ville Hautamäki, Nicholas Evans, and Zheng-Hua Tan, "Utterance Verification for Text-Dependent Speaker Recognition: A Comparative Assessment Using the RedDots Corpus," in *Proc. Interspeech 2016*, 2016, pp. 430–434.

[8] Gilles Boulianne, "Language-independent voice passphrase verification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4490–4494.

[9] Hossein Zeinali, Hossein Sameti, and Themos Stafylakis, "DeepMine Speech Processing Database: Text-Dependent and Independent Speaker Verification and Speech Recognition in Persian and English ," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2018)*, 2018, pp. 386–392.

[10] Ivx00E1;n Lx00F3;pez-Espejo, Zheng-Hua Tan, John H. L. Hansen, and Jesper Jensen, "Deep spoken keyword spotting: An overview," *IEEE Access*, vol. 10, pp. 4169–4199, 2022.

[11] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *CoRR*, vol. abs/1804.03209, 2018.

[12] Raphael Tang, Jaejun Lee, Afsaneh Razi, Julia Cambre, Ian Bicking, Jofish Kaye, and Jimmy Lin, "Howl: A deployed, open-source wake word detection system," in *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, Eunjeong L. Park, Masato Hagiwara, Dmitrijs Milajevs, Nelson F. Liu, Geeticka Chauhan, and Liling Tan, Eds., Online, Nov. 2020, pp. 61–65, Association for Computational Linguistics.

[13] Yiming Wang, Hang Lv, Daniel Povey, Lei Xie, and Sanjeev Khudanpur, "Wake word detection with streaming transformers," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5864–5868.

[14] Zhongxin Bai and Xiao-Lei Zhang, "Speaker recognition based on deep learning: An overview," *Neural Networks*, vol. 140, pp. 65–99, 2021.

[15] Jinlong Xue, Yayue Deng, Yichen Han, Ya Li, Jianqing Sun, and Jiaen Liang, "Ecapa-tdnn for multi-speaker text-to-speech synthesis," in *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 2022, pp. 230–234.

[16] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-François Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech Language*, vol. 64, pp. 101114, 2020.

[17] Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[18] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694.

[19] Tianchi Liu, Rohan Kumar Das, Maulik Madhavi, Shengmei Shen, and Haizhou Li, "Speaker-Utterance Dual Attention for Speaker and Utterance Verification," in *Proc. Interspeech 2020*, 2020, pp. 4293–4297.

[20] Yan Liu, Zheng Li, Lin Li, and Qingyang Hong, "Phoneme-Aware and Channel-Wise Attentive Learning for Text Dependent Speaker Verification," in *Proc. Interspeech 2021*, 2021, pp. 101–105.

[21] Hossein Zeinali, Kong Aik Lee, Jahangir Alam, and Lukáš Burget, "SdSV Challenge 2020: Large-Scale Evaluation of Short-Duration Speaker Verification," in *Proc. Interspeech 2020*, 2020, pp. 731–735.

[22] G. Sell D. Povey S. Khudanpur D. Snyder, D. Garcia-Romero, "Musan: A music, speech, and noise corpus," 2015.

[23] Gilles Boulianne Lukas Burget Ondrej Glembek Nagendra Goel Mirko Hannemann Petr Motlicek Yanmin Qian Petr Schwarz Jan Silovsky Georg Stemmer Karel Vesely Daniel Povey, Arnab Ghoshal, "The kaldi speech recognition toolkit," in *IEEE Signal Processing Society*, 2011.

[24] Hossein Zeinali, Kong Aik Lee, Jahangir Alam, and Lukas Burget, "Short-duration speaker verification (sdsv) challenge 2021: the challenge evaluation plan," 2021.

[25] Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *ArXiv*, vol. abs/2006.11477, 2020.

[26] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.