INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Segment-level Effects of Gender, Nationality and Emotion Information on Text-independent Speaker Verification

Kai Li[1], *Masato Akagi*[1], *Yibo Wu*[2], *Jianwu Dang*[2]

[1]Japan Advanced Institute of Science and Technology, Japan
[2]Tianjin University, China

kai_li@jaist.ac.jp, akagi@jaist.ac.jp, yibo_wu@tju.edu.cn, jdang@jaist.ac.jp

Abstract

Speaker embeddings extracted from neural network (NN) achieve excellent performance on general speaker verification (SV) missions. Most current SV systems use only speaker labels. Therefore, the interaction between different types of domain information decrease the prediction accuracy of SV. To overcome this weakness and improve SV performance, four effective SV systems were proposed by using gender, nationality, and emotion information to add more constraints in the NN training stage. More specifically, multitask learning-based systems which including multitask gender (MTG), multitask nationality (MTN) and multitask gender and nationality (MTGN) were used to enhance gender and nationality information learning. Domain adversarial training-based system which including emotion domain adversarial training (EDAT) was used to suppress different emotions information learning. Experimental results indicate that encouraging gender and nationality information and suppressing emotion information learning improve the performance of SV. In the end, our proposed systems achieved 16.4 and 22.9% relative improvements in the equal error rate for MTL- and DAT-based systems, respectively.

Index Terms: Multitask learning, Domain adversarial training, Speaker embedding, Text-independent speaker verification

1. Introduction

The purpose of a speaker verification (SV) system is to verify whether two test utterances belong to the same speaker based on the speaker characteristics extracted from the raw speech. Speech is a complex signal that conveys many types of information such as linguistic content, speaker individuality, nationality, gender, and emotion. Some information may be useful for SV, while others may not. Therefore, the performance of SV could be increased by enhancing meaningful information and suppressing useless information.

Speaker characteristics extracted from deep neural networks (DNNs) [1] [2] [3] [4] [5] [6], which are known as speaker embeddings, are used in SV and have been used to develop state-of-the-art methods. However, the learning ability of a DNN is still limited because of the practice of using speaker labels only in the training stage of SV system, which does not take into account the interaction between different types of domains information.

Multitask learning (MTL)[7][8], recently proposed to learn useful information for SV by using speaker-irrelevant labels, and domain adversarial training (DAT)[9][10], designed to eliminate the effect of useless information by using a gradient reversal layer (GRL) in different domains, have significantly improved the performance of SV. What these two approaches have in common is that they add more constraints during the neural network (NN) training stage. For example, phonetic information was chosen as an aid in speaker recognition by using MTL[11]. The results from that study suggested that phonetic information is useful for speaker recognition at the frame level. The results from Wang et al.[12] also showed that extracting phonetic information at the segment level decreases the final performance of SV, while suppressing such information by using DAT can further improve performance. Zhou Meng et al. [13] improved the performance of SV by using DAT to suppress the learning of the environment and signal-to-noise-ratio variability information. Tu et al.[14] showed that the performance of speech emotion recognition (SER) degrades because of the effect of speaker information. However, gender and nationality information are crucial in verifying the identity of a speaker because these information can be used as multiple verifications. Utterances uttered by the same speaker in different emotions vary greatly in their characteristics which also influence the extraction of speaker individual features and decrease the accuracy.

Subjectively, gender and nationality are speaker-invariant information. This means that, for a given speaker in a training database, they will not change and can provide additional information for the authentication of speaker identity. Therefore, these two information should be beneficial for SV. On the contrary, emotion information can change in different speaking scenarios, which will decrease the cosine similarity score in test pairs even though the two utterances are from the same speaker; therefore, it has to be suppressed.

Based on these reasons, we investigated the effects of gender, nationality, and emotion information on the performance of SV systems. Four effective systems were proposed by using MTL- and DAT-based methods. MTL-based systems which including multitask gender (MTG), multitask nationality (MTN) and multitask gender and nationality (MTGN) were used to enhance gender and nationality information learning in the NN training stage. DAT-based system which including emotion domain adversarial training (EDAT) was used to suppress different emotions information learning.

The rest of this paper is organized as follows. In Section 2, we describe the MTL- and DAT-based NN architectures of the proposed system. We then introduce the databases and detail configuration of our experiments in Section 3. We report the experimental results and provide discussion in Section 4. Finally, we conclude the paper in Section 5.

2. Proposed methods

Our purpose was to extract segment-level embeddings from NNs that contain as much speaker-invariant information as possible. In this section, we introduce the MTL- and DAT-based NN architectures of the proposed systems for gender,

http://dx.doi.org/10.21437/Interspeech.2020-1700

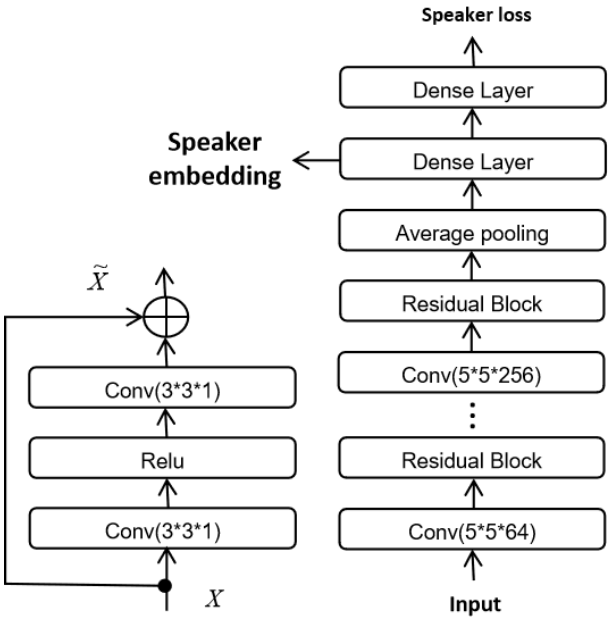

Figure 1: *Architecture of residual-network-based speaker-embedding extractor*

nationality, and emotion information. For convenience, the data that fed into NN in the training stage are represented as $\{(X_i, g_i, n_i, e_i)\}_{i=0}^{N}$, where $X_i = [x_i^1, x_i^2, ..., x_i^t] \epsilon R^{t \times d}$ refers to the front-end input of utterance i with fixed number of frames t and the feature dimensions of each frame are d. The notations g_i, n_i, e_i refer to labels of gender, nationality, and emotion information, respectively. N is the total number of training utterances.

2.1. Residual network (ResNet)-based speaker embeddings extraction

As computing power and storage capacity increases, NNs become deeper and deeper[15][16]. However, it becomes very difficult to achieve a stable performance in NN training stage because of the gradient vanishing problem. To solve this problem, ResNet was proposed[17] and used to develop state-of-the-art methods[18][19] for SV systems. In this study, we constructed a ResNet-based SV system as our baseline system to compare it with the proposed system.

The architecture of this ResNet-based SV system is shown in Figure 1. Three one-dimensional convolutional layers combined with three residual blocks are used to generate a frame-level feature for utterance X_i. For the three convolutional layers, the kernel size is (5×5) and the number of channels varies from 64 to 256. For each residual block, two convolution layers with the same kernel size (3×3) and stride (1×1) and a rectified-linear-unit function are used in the back of the first convolution layer. After the average pooling layer, segment-level speaker embeddings can be extracted from a 1024-dimentional fully connected layer (FCL). Then, the embeddings are mapped into a number corresponding to that of speakers in the training data. Finally, angular softmax (A-softmax) loss[20] is used to optimize the entire network.

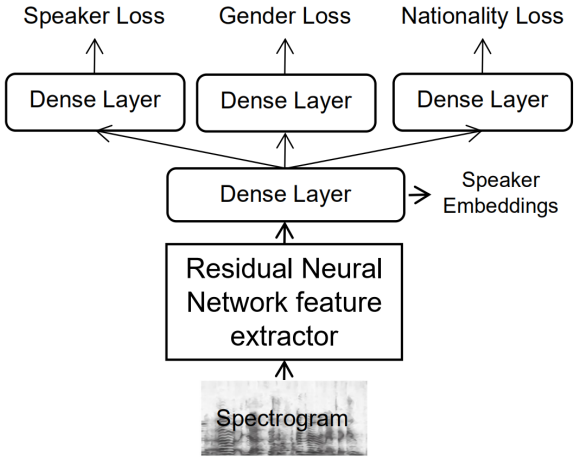

Figure 2: *Multitask learning (MTL)-based architecture of proposed system by using information of gender and nationality*

2.2. Multitask learning for gender and nationality information

To achieve multiple verification of gender and nationality information, MTL-based NN architectures (MTG, MTN and MTGN) that includes one or two more sub-branches to classify the labels of gender and nationality information is illustrated in Figure 2. This architecture is based on the ResNet architecture described in 2.1. Speaker embeddings e_i, extracted from the ResNet-based feature extractor, are fed into two FCLs that have the same number of dimensions as gender and nationality labels. Then, cross-entropy loss is used to calculate the discrepancy between the embedding features and attribute labels. To minimize the loss of different domains simultaneously, the final objective functions we used in this study are

$$L_{MTG} = L_A^S + \alpha L_{CE}^G \tag{1}$$

$$L_{MTN} = L_A^S + \beta L_{CE}^N \tag{2}$$

$$L_{MTGN} = L_A^S + \alpha L_{CE}^G + \beta L_{CE}^N \tag{3}$$

where L_A^S is the loss of the speaker classifier by using A-Softmax loss. The notations L_{CE}^G and L_{CE}^N refer to the cross-entropy loss for gender and nationality classifiers, respectively. We use α and β as trade-offs to control the effects of gender and nationality information on SV, respectively. To reach the best performance in MTL-based architectures of we proposed, different trade-offs α/β are rivised in the training stage. The results of all the experiments are disscussed in Section 4.

2.3. Domain adversarial training with emotion information

We also constructed a DAT-based NN architecture (EDAT), as shown in Figure 3, which is also based on the ResNet architecture described in 2.1. The purpose of DAT is to minimize the divergences between the emotion and speaker domains so that we can extract domain-invariant and speaker-discriminative features. As depicted in Figure 3, the only difference between MTL and DAT is that a GRL is used with DAT to inverse the gradient flow from the emotion domain. The total loss of DAT can be defined as

$$L_{EDAT} = L_A^S - \lambda L_{CE}^E \tag{4}$$

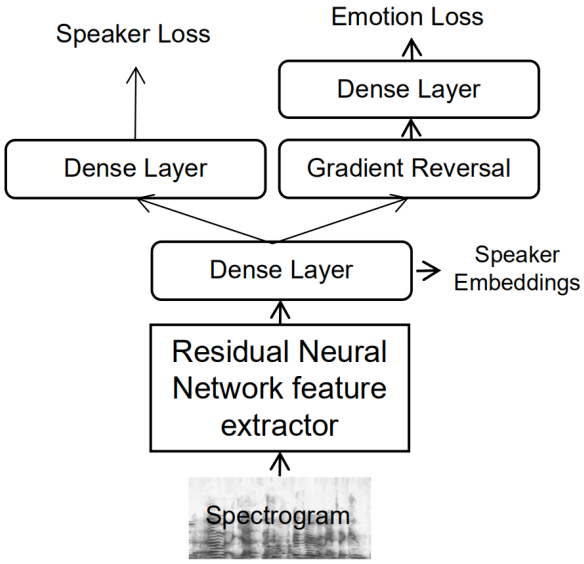

Figure 3: *Domain adversarial training (DAT)-based NN architecture of proposed system by using labels of emotion*

where L_{CE}^{E} is the cross-entropy loss for the emotion classifier. These two classifiers are trained with an adversarial purpose. The model parameters of speaker classification are adjusted to minimize the L_{A}^{S}, and those of emotion classification are adjusted to maximize the L_{CE}^{E}. The minimax competition enhances the discrimination of different speakers, suppresses the discrimination of different emotions, and converges to a point where the embeddings we extracted cannot be correctly classified into different emotions. Therefore, the embeddings we obtained will not be affected by emotion information under ideal conditions.

2.4. Score fusion and calibration

To fuse multiple subsystems that we introduced above into a single system, a score fusion technique [21] which based on linear regression algorithm is used for score fusion and calibration. The purpose of this technique is to construct a mapping ξ that maps scores from different subsystems into a log-likelihood-ratio. The mapping can be depicted by the following formula:

$$\xi_t = a + \sum_{1}^{n} b_i s_{it} + q_t^{'} W r_t \tag{5}$$

where ξ_t is the fused log-likelihood-ratio for a test pair t, N is the number of fused subsystems, s_{it} is the score of subsystem i for trial t, and q_t and r_t are optional quality vectors. Scalar offset a, scalar combination weights b_i, and symmetric matrix W should be optimized based on the scores of the development and test files. The test pairs of the development file are randomly generated from 1211 training speakers. This file includes approximately 40000 test pairs. The test file use the original Voxceleb1 test file described later. After 100 iterations, ξ reaches superior performance.

3. Experimental setup

3.1. Dataset

To verify the effects of gender, nationality, and emotion information on SV, We used the Voxceleb database, which includes gender and nationality labels, and IEMOCAP database, which is very popular in SER and provides emotion labels.

3.1.1. Voxceleb

There are two versions of the Voxceleb database[22], and we just used the first version (Voxceleb1). All the utterances were extracted from videos uploaded to YouTube and encoded at 16 kHz sampling rate. The database is gender balanced and includes candidates from approximately 35 nationalities. The details of the training and test sets of this database are summarized in Table 1.

Table 1: *Training and test sets statistics of Voxceleb1 database*

Set	Speaker	Utterance
Train	1,211	148,642
Test	40	4,874
Total	1,251	153,516

3.1.2. IEMOCAP

IEMOCAP[23] is a popular database in SER that includes approximately 12 hours of audio data from 10 speakers. These audio data were recorded from dialogues between a male and female on different topics then divided into utterances of 3 to 15 s. Ten emotional labels were manually labeled but only four (anger, excitement, neutral and sadness) including 5531 utterances were selected to investigate the effect of emotion on SV.

3.2. Experimental conditions

In the training stage, 3-s utterances are randomly selected from each raw waveform. Then a 161-dimensional spectrogram is extracted from each frame with a frame length of 20 ms and frame shift of 10 ms. 128 utterances are grouped as one batch fed into different speaker-embedding-extraction systems, and the training epochs of each system are set to 80. The detailed configurations of the architecture of ResNet-based SV systems are described in Section 2. In the test stage, the whole waveform is used to extract speaker embeddings to obtain more reliable scores. The equal error rates and minimum decision cost function, introduced in national institute of standards and technology speaker recognition evaluation plan 2012[24], were used as the evaluation index in our study. Voxceleb1 was used only for training of MTL-based NN architectures and IEMOCAP was used for training of DAT-based NN architecture. All the results of MLT-based systems were evaluated on Voxceleb1, which has 37720 test pairs from 40 speakers. A 5-fold cross-validation method with approximately 20000 test pairs generated from two speakers for each time validation was used for DAT-based system to evaluate the performance of the proposed system.

4. Results and discussion

The results in Figure 4 indicate that the effects of gender and nationality information in MTG and MTN systems differ and reach the best perpormance when α and β are set to 1 and 0.7

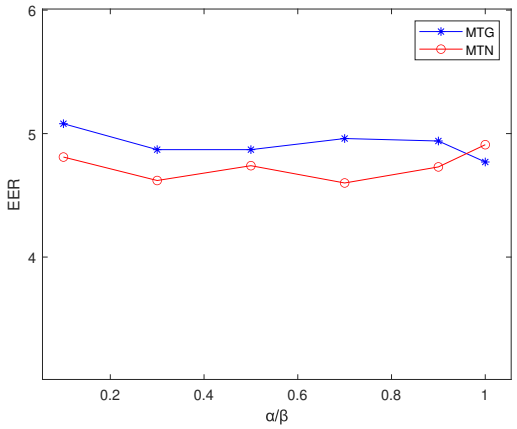

Figure 4: *Difference of gender and nationality information in equal error rates (EERs) by revising different trade-offs α/β*

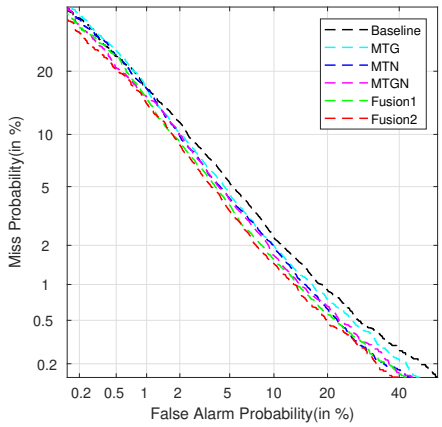

Figure 5: *Detection error tradeoff (DET) curves of proposed gender/nationality/emotion-aware systems and ResNet baseline system*

respectively. However, the MTGN system, could not further improve in performance when α and β were set to 1 and 0.7, respectively. This is because gender and nationality information affect each other during the training stage. Finally, the system can improve in performance when α and β are set to 0.1 at the same time. The results from MTG, MTN and MTGN are listed in Table 2.

Table 2: *Results of MTL- and DAT-based systems in terms of EER and minimum decision cost function (minDCF) based on two different databases*

System	Database	EER(%)	minDCF
ResNet Baseline	Voxceleb1	5.12	0.051
MTG	Voxceleb1	4.77	0.047
MTN	Voxceleb1	4.60	0.046
MTGN	Voxceleb1	4.59	0.045
Fusion1	Voxceleb1	4.40	0.044
Fusion2	Voxceleb1	4.28	0.042
ResNet Baseline	IEMOCAP	16.78	0.165
EDAT	IEMOCAP	12.94	0.128

The results of MTL-based systems (MTG, MTN, MTGN, fusion1 and fusion2) and results of DAT-based system (EDAT) regarding gender, nationality and emotion are listed in Table 2. The results indicate that gender and nationality information can improve the performance of SV. To further improve SV performance, a score-linear-fusion toolkit that introduced in Section 2[21] was used to carry out the score fusion. Fusion1 denotes the score fusions of the MTG and MTN systems, and Fusion2 denotes the score fusions of the MTG, MTN, and MTGN systems. The EER of the fusion2 improved with an EER of 16.4% over the baseline system with an EER of 5.12%. Also, the performance of SV was affected by emotion information and can be improved by suppressing the learning of emotion information. The EDAT system reduced the EER of the baseline system (16.78%) to 12.94%, a relative improvement of 22.9%.

Figure 5 shows the comparison of detection error tradeoff curves[25] between the proposed system and baseline system. Finally, the fusion of different proposed systems performs the best.

5. Summary

We investigated the effects of gender, nationality and emotion information on SV by using MTL and DAT. Four effective systems were proposed by combining speaker information and different types of domain information in NN training stage. More specifically, MTL-based method was used to enhance the learning of gender and nationality information in MTG, MTN and MTGN systems. The information learning of different emotions of a certain speaker was suppressed using DAT-based method in EDAT system. Finally, a linear scoring fusion method was employed to combine the advantages of different systems. The results indicate that enhance gender and nationality information learning by using MTL-based methods can significantly improve the performance of SV. The results also indicate that the effect of emotion information is suppressed by using DAT-based method also beneficial for SV. Finally, compared with a baseline system, the performance of our systems achieved 16.4% and 22.9% relative improvements in the EER of MTL and DAT-based systems, respectively. Moreover, the relationship of different speech information can also be referenced to improve the recognition performance in other research fields such as SER and nationality recognition.

6. References

[1] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4052–4056.

[2] D. Snyder, P. Ghahremani, D. Povey, D. Garcia-Romero, Y. Carmiel, and S. Khudanpur, "Deep neural network-based speaker embeddings for end-to-end speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 165–170.

[3] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[4] L. Li, Y. Chen, Y. Shi, Z. Tang, and D. Wang, "Deep speaker feature learning for text-independent speaker verification," *arXiv preprint arXiv:1705.03670*, 2017.

[5] G. Bhattacharya, M. J. Alam, and P. Kenny, "Deep speaker embeddings for short-duration speaker verification." in *Interspeech*, 2017, pp. 1517–1521.

[6] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5115–5119.

[7] A. H. Poorjam, M. H. Bahari *et al.*, "Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals," in *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*. IEEE, 2014, pp. 7–12.

[8] Z. Tang, L. Li, D. Wang, R. Vipperla, Z. Tang, L. Li, D. Wang, and R. Vipperla, "Collaborative joint training with multitask recurrent model for speech and speaker recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 3, pp. 493–504, 2017.

[9] Z. Meng, J. Li, Z. Chen, Y. Zhao, V. Mazalov, Y. Gang, and B.-H. Juang, "Speaker-invariant training via adversarial learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5969–5973.

[10] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4889–4893.

[11] Y. Liu, L. He, J. Liu, and M. T. Johnson, "Speaker embedding extraction with phonetic information," *arXiv preprint arXiv:1804.04862*, 2018.

[12] S. Wang, J. Rohdin, L. Burget, O. Plchot, Y. Qian, K. Yu, and J. Černockỳ, "On the usage of phonetic information for text-independent speaker embedding extraction," *Proc. Interspeech 2019*, pp. 1148–1152, 2019.

[13] Z. Meng, Y. Zhao, J. Li, and Y. Gong, "Adversarial speaker verification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6216–6220.

[14] M. Tu, Y. Tang, J. Huang, X. He, and B. Zhou, "Towards adversarial learning of speaker-invariant representation for speech emotion recognition," *arXiv preprint arXiv:1903.09606*, 2019.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018.

[19] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, 2017.

[20] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[21] N. Brümmer and E. De Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865*, 2013.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[24] C. S. Greenberg, "The nist year 2012 speaker recognition evaluation plan," *NIST, Technical report*, 2012.

[25] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," National Inst of Standards and Technology Gaithersburg MD, Tech. Rep., 1997.