INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Whisper to neutral mapping using cosine similarity maximization in i-vector space for speaker verification

Abinay Reddy Naini, Achuth Rao MV, Prasanta Kumar Ghosh
Electrical Engineering, Indian Institute of Science, Bangalore 560012, India
nainireddy@iisc.ac.in, achuthr@iisc.ac.in, prasantg@iisc.ac.in

Abstract

In this work, we propose a novel feature mapping (FM) from whispered to neutral speech features using a cosine similarity based objective function for speaker verification (SV) using whispered speech. Typically the performance of an SV system enrolled with neutral speech degrades significantly when tested using whispered speech, due to the differences between spectral characteristics of neutral and whispered speech. We hypothesize that FM from whispered Mel frequency cepstral coefficients (MFCC) to neutral MFCC by maximizing cosine similarity between neutral and whisper i-vectors yields better performance than the baseline method, which typically performs a direct FM between MFCC features by minimizing mean squared error (MSE). We also explored an affine transform between MFCC features using the proposed objective function. Whisper SV experiments with 1882 speakers reveal that the equal error rate (EER) using the proposed method is lower than that using the best baseline by $\sim$24% (relative). We show that the proposed FM system maintains the neutral SV performance, while improving the EER of whispered SV unlike baseline methods. We also show that the bias in the learned affine transform is corresponds to the glottal flow information, which is absent in the whispered speech.

Index Terms: Speaker verification, whispered speech, feature mapping, cosine similarity.

1. Introduction

Over the past two decades, speaker verification (SV) systems gained a lot of attention due to its wide applications in the field of speech forensics and biometrics [1]. A typical SV system is trained to verify whether a test speech sample belongs to any of the enrolled speakers [2]. Unlike in speaker recognition, where it always maps the test speech to one of the closest enrolled speakers, an SV system has to reject if the test speech is from an imposter [2].

Extensive research has been done on SV systems to make them robust to noise conditions, but less emphasis has been given for varying vocal effort like whispered speech [3]. Whisper is one of the natural modes of speech production, and it is often used by speakers in their day-to-day life. For example, speakers often whisper while sharing credit card information and passwords to a biometric system to avoid the content being overheard [4]. In some cases, criminals might whisper in a telephonic conversation to avoid sharing voice print with the forensic authorities [5]. For some people, such as laryngectomees, whisper is the only mode of speech production [6]. These bring up a necessity for making an SV system robust to whispered speech, without compromising the performance of the neutral speech based SV.

The speech spectrum of whispered speech is significantly different from that of the neutral speech. The major differences include the absence of voicing [7], flat spectral slope [8, 4], and a major shift in low-frequency formants [9]. It is also shown that speakers, while whispering, hyper-articulate to ensure intelligibility [10]. Despite all these differences, previous works in the literature showed that whispered speech contains adequate information about speech, gender, and the speaker [7]. However, it is challenging to improve SV system performance when it is enrolled with neutral speech and tested with whispered speech. This is mainly due to the absence of pitch [4] and a shift in lower frequency formants [9].

SV using neutral speech has been well studied in recent years. Among these works, Front-end factor analysis (i-vector) [11] and Deep Neural Network (DNN) embedding based methods [12, 13] are considered to be the state-of-the-art neutral speech based SV methods. The detailed review of neutral SV methods can be found in [14, 15]. Unlike neutral SV, in whispered SV, enrollment is done using neutral and/or whispered speech and only whispered speech is used during the test. Despite the limited research on whispered SV, there exist few attempts to reduce the gap between whispered SV and neutral SV. Recently, the DNN embedding based methods showed substantial improvements in neutral SV [13]. But extending these methods to whispered SV is difficult due to non-availability of large whispered speech corpus. Whispered SV methods can be broadly classified into two categories. In the first category, authors explored different features [16, 17, 18, 19, 20] to improve the whispered SV performance. Among these features, some have been used along with Gaussian mixture models (GMM) [21] for SV task, others are applied with the front end as the i-vector [11]. In the second category, the authors explored different feature transformation methods such as frequency warping [8] and DNN based feature mapping (FM) [22]. A detailed survey of different whispered SV methods can be found in [23]. In DNN based FM [22], from whispered to neutral features, authors explored the mel frequency cepstral coefficients (MFCC). But the authors trained the FM by optimizing the mean squared error (MSE) between the predicted MFCC and the dynamic time warped (DTW) neutral MFCC. In the context of FM for whispered SV, the main goal is to map the features from whispered to neutral speech by retaining only the speaker information. But MSE doesn't differentiate between the speaker-specific factors. Hence, we hypothesize that the MSE based objective function is not appropriate in the context of whispered SV. Rather, we believe, that a mapping that is targeted to preserve speaker-specific information would be more useful.

In this paper, we propose a novel objective function based on cosine similarity in i-vector space for learning a FM from whispered to neutral speech. We use this mapping with the front end i-vector setup to perform whispered SV. Whispered SV experiments with 1882 speakers comprising 186708 neutral and 26892 whispered recordings reveal that the equal error rate (EER) using the proposed method is lower than that using the best baseline by $\sim$24% (relative). We show that the affine transform based FM performs better than other FM techniques.

http://dx.doi.org/10.21437/Interspeech.2019-2280

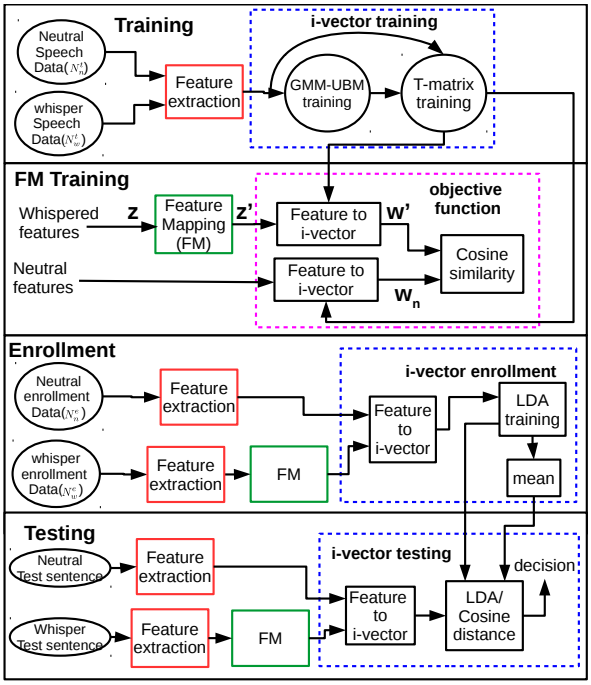

Figure 1: *Block diagram of the whispered speaker verification system with the proposed feature mapping (FM). Feature extraction step is shown in red, Front end i-vector step is shown in blue , FM is shown in green and proposed objective function is in magenta.*

2. Proposed whispered SV system

The block diagram of the proposed whispered SV is shown in Fig. 1. All the blocks are explained in detail below.

2.1. Feature extraction

MFCCs are widely used features in different speech applications. At first, a given speech signal is pre-emphasized with a filter coefficient of α. The pre-emphasized signal is used to obtain a 13-dimensional MFCC feature vector. These features are computed over a window length T_w with a shift of T_s. To add temporal dynamics to the feature vector, velocity and acceleration coefficients are computed resulting in a 39-dimensional feature vector [18].

2.2. Front end i-vector

Front end i-vector setup involves three main steps as follows.

i-vector training: Feature vectors of speech can be modeled as speaker adapted GMM from Universal background model (UBM). The parameters of the UBM is indicated by Ω. The mean vectors in the GMMs are concatenated to construct a super vector ($\mathbf{m}_s$) of dimension CF, where C is the number of mixtures and F is the dimension of the feature vector. In i-vector, the super vector is modeled using $\mathbf{m}_s = \mathbf{m} + Tw$, where is T is a tall and low rank matrix of dimension $CF \times d$ and $\mathbf{m}$ is a speaker and channel independent super vector and w is referred as the i-vector (with dimension d). Given the training data and the UBM, the matrix T is learned using the Expectation-Maximization algorithm [24].

Feature to i-vector: Given a sequence of features $\mathbf{y} = [y_1, y_2, \ldots, y_L]$ of length L and dimention F. The i-vector is computed as follows [11]:

$$w = (I + T^t \Sigma^{-1} N(y) T)^{-1} T^t \Sigma^{-1} \hat{F}(\mathbf{y}) \quad (1)$$

where $I_{d\times d}$ is an identitiy matrix, $N(\mathbf{y})_{CF\times CF}$ is a block diagonal matrix whose diagonal entries are $N_c I_{F\times F}, 1 \leq c \leq C$. $\hat{F}(\mathbf{y})$ is a supervector of dimention CF obtained by concatenating first-order BaumWelch statistics $\hat{F}_c, 1 \leq c \leq C$ for a given utterance ($\mathbf{y}$). Σ is a diagonal covaiance matrix of error in $\mathbf{m}_s$ modeling, which is estimated during the training. $\hat{F}_c$ and N_c are computed as follows: $N_c = \sum_{t=1}^{L} p(c|y_t, \Omega)$, $\hat{F}_c = \sum_{t=1}^{L} p(c|y_t, \Omega)(y_t - m_c)$, where m_c is the mean vector of the c-th UBM mixture [11].

i-vector enrollment/testing: In the enrollment step, we extract i-vectors for each neutral and whispered speech features as discusssed above, and by taking an average of all these i-vectors of a speaker we obtain one final reference i-vector for each speaker (i^{ref}). We also learn a linear discriminant analysis (LDA) for the set of enrollment speakers to find the subspace where the enrollment speakers are linearly discriminated.

During the test, we first compute i-vector for a test recording (i^{test}) and reduce the dimension of both i^{ref} and i^{test} using the learned LDA to get l^{ref} and l^{test} respectively. The cosine similarity (ϕ) between these vectors is computed as follows:

$$\phi(l^{ref}, l^{test}) = \frac{< l^{ref}, l^{test} >}{||l^{ref}||\ ||l^{test}||}, \quad (2)$$

where $< a, b >$ indicates the inner product between the vectors a and b and $||a||$ indicate the norm of the vector a. We also refer $\psi = (1 - \phi)$ as the cosine distance. Finally, an appropriate threshold is applied on the cosine distance to accept/reject the speaker.

2.3. Feature Mapping (FM)

Typically, the distribution of MFCC for whispered speech is different from that of the neutral speech. Hence, DNN based FM is used to map from whispered feature ($\mathbf{x}$) to neural features ($\mathbf{y}$) [22]. We also explore an affine transformation based FM as follows $\mathbf{y} = A\mathbf{x} + b$. Sarria et al. [22] used an MSE based objective function to learn the FM using DTW aligned whispered and neural features. But, we hypothesize that the MSE may not be the right objective function for the i-vector based SV because the MSE doesn't guarantee to retain only the speaker information after mapping.

2.4. Proposed objective function

In the enrollment stage of whispered SV, the i-vectors are computed using only neutral speech. The main goal of the proposed FM is to transform the feature from whispered speech such that the i-vector computed from the transformed feature is closer to the neural i-vector in terms of cosine similarity. Hence, we propose to learn FM using an objective function which maximizes the cosine similarity in the i-vector space.

Given whispered feature vectors for an utterance $\mathbf{z} = [z_1, z_2 \ldots, z_L]$, is mapped to the features $\mathbf{z}' = [z_1', z_2' \ldots, z_L']$ using a mapping f_θ, where f_θ can be any FM function discussed above with the parameter set $\{\theta_i, 1 \leq i \leq N\}$. From $\mathbf{z}'$ the i-vectors are computed using eq. 1 and it is denoted by w'. Similarly from the neutral features, the computed i-vectors are denoted by w_n (as shown in Fig 1). In the FM training, the information about the LDA matrix is not available and it depends

on the enrollment speakers. Hence, we propose to maximize the cosine similarity between the neutral i-vector w_n and the i-vector of the mapped features $w^{'}$. The cosine similarity between w_n and $w^{'}$ is indicated by $\phi(w_n, w^{'})$ [11]. To train the FM using the cosine similarity, the derivative of $\phi(w_n, w^{'})$ with respect to mapping parameters are derived as follows:

$$\frac{d\phi(w_n, w^{'})}{d\theta_i} = \left[\frac{d\phi(w_n, w^{'})}{dw^{'}}\right]^T \frac{dw^{'}}{dz^{'}} \frac{dz^{'}}{d\theta_i},\ 1 \leq i \leq N \quad (3)$$

The term $\frac{d\phi(w_n, w^{'})}{dw^{'}}$ is computed from eq. 2, $\frac{dw^{'}}{dz^{'}}$ is computed from eq. 1 and $\frac{dz^{'}}{d\theta_i}$ can be derived using the type of mapping function. Note that the proposed method doesn't require DTW alignment because we compute one i-vector per utterance and maximize the cosine similarity in the i-vector space.

3. Experiments and results

3.1. Database

We consider 5 different databases for this study: (1) CHAINS corpus [25] contains 36 speakers among them 8 are from the UK and the USA, remaining 28 are from the Eastern part of Ireland. The CHAINS data was recorded in 6 different modes with varying vocal effort, among which we consider only whisper and solo neutral speech from all 36 speakers recorded at 44.1kHz. (2) wTIMIT (whisper TIMIT) [26] contains 48 speaker's recordings of 460 MOCHA-TIMIT sentences [27], recorded in both neutral and whispered mode at 44.1kHz. Among these 48 speakers, 28 are from northern America and the remaining 20 are from the Singapore. (3) VoxCeleb1 [28] is a large audio-visual dataset consisting of short audio clips, which were extracted from YouTube interview videos of celebrities from all over the world. We consider a total of 1251 speakers data, which was recorded at 16kHz. (4) TIMIT data set [29] contains a total of 630 speakers from eight major dialect regions of the USA with only 10 neutral sentences per speaker, which were recorded at 16kHz. (5) We consider an in-house recorded data containing 9 speakers from two dialect regions of India. Each speaker speaks 460 MOCHA-TIMIT sentences [27] in both neutral and whispered mode recorded at 16kHz. Details of this in-house dataset can be found in [30], which is refered as wSPIRE in this paper. We resample recordings from all corpora to a common sampling frequency of 16kHz.

3.2. Experimental setup

In the Experimental stage, we divide recordings from all five datasets into training and testing speakers. The training and test splits of the databases are provided in Table 1. Unlike, UBM, T matrix and FM training, not all utterances for a speaker are used for enrollment/LDA training and testing. In particular, only 10 neutral/whisper utterance pairs are chosen for this purpose except for TIMIT, where only 10 neutral utterances are chosen from selected 100 speakers. Among these 10 utterances, 8 are used for enrollment/LDA training and remaining two neutral/whisper pair from wTIMIT and CHAINS are used for testing. We consider two experiments. In the first experiment, eight neutral utterances ($N_n^e = 8$) are used for enrollment and no whisper data is used ($N_w^e = 0$). In the second experiment, we vary the number of whisper utterances $N_w^e = 1, 2, 4, 6, 8$ along with eight neutral utterances for enrollment.

We extract MFCC features using $\alpha = 0.97, T_w = 25ms$ and $T_s = 10ms$. For extracting i-vectors we use $C = 512, F = 39$ and $d = 400$. In the DNN based feature mapping we use 2 layers of DNN with 256 and 128 units with relu activation. In the affine (Af) transform based FM, we consider two cases- 1) unconstrained A and 2) A constrained to be an identity matrix.

Table 1: *Number of male/female speakers and recordings per speaker for all three databases considered in this work. $*$ indicates that the number can be different depeding on the experimental condtion. tr indicates training.*

Data split	Num. of Speakers/database					Total Recordings	
	VoxCeleb1	wTIMIT	TIMIT	CHAINS	wSPIRE	Ne	Wh
UBM tr	1251	0	462	0	0	157k	0
T matrix tr	0	24	462	0	9	19.4k	14.8k
FM tr	0	14	0	0	9	14.8k	14.8k
Enrollment/ LDA tr	0	24	100	36	0	1280	480*
Testing	0	24	100*	36	0	320*	320*
# of FEMALE	563	16	192	16	3	-	-
# of MALE	688	20	438	20	6	-	-

We consider two baseline schemes. In the first baseline, no FM is performed and i-vectors are computed directly using the MFCCs of the test utterance. We refer this method as WFM. In the second baseline (FM_{MSE}), we transform whisper MFCCs using DNN/Affine based FM. This FM is trained by optimizing MSE between the DTW aligned whisper and neutral MFCCs. In the proposed method (FM_{cs}), we transform whisper MFCCs using DNN/Affine based FM, but the mapping is trained using the proposed objective function.

After training the FM, it can be used in the verification system in two ways. The first way is to directly use FM in the enrollment stage to transform the whispered feature and compute the i-vectors. But due to the feature transformation, there could be mismatch between the supervector corresponding to the i-vector computed from these transformed features and the subspace spanned by the training supervectors. Hence in the second method, we retrain the T-matrix using the neural features and the mapped whispered features. The re-trained T-matrix and mapping are used in both enrollment and test stages.

To understand the effect of type of the data used for T-matrix training, we consider two types of data. The first type (Ne) uses features computed only using neutral data and the second type (Ne+Wh) uses features computed using both neutral and whisper data.

We use equal error rate (EER) as an evaluation metric for SV, which is the error rate of SV system when the false acceptance rate of the imposter and the false rejection rate of the enrolled speakers are equal [31]. We have implemented the feature mapping in tensorflow [32] and keras [33]. We have optimized the objective function using the gradients shown in eq. 3 and adam optimizer [34], until the validation error increases.

3.3. Results & discussion

Table 2 shows, the comparison of EER for different methods in different experimental conditions as discussed in the previous section (Some combinations are omitted because of poor EER). It can be observed from the table that, in all cases, EER_{ne} is considerably better than EER_{wh}. It indicates that the i-vectors computed using only neutral data deviates significantly from i-vectors computed using only whisper data in terms of cosine similarity. It is also clear from the table that, including the whisper data in T matrix training improves the EER_{wh}. But it results in a slight degradation of EER_{ne}. In FM_{MSE}, it can be observed that, retraining the T-matrix after FM improves the EER_{wh} (from 23.7 to 20.2). But this causes a slight increase in EER_{ne}. It can be observed from Table 2 that, Af based FM is

Table 2: *Performance comparison of proposed method and baseline methods for different experimental conditions with $N_w^e = 0$. EER_{ne} and EER_{wh} indcate the EER for only neutral and only whispered test utterances respectively.*

Method	WFM		FM_{MSE}					FM_{cs}			
mapping type	-		DNN		Af			DNN	Af		Af $(A = I)$
T data	Ne	Ne+Wh	Ne+Wh	Ne+Wh	Ne+Wh	Ne	Ne+Wh	Ne+Wh	Ne	Ne+Wh	Ne+Wh
T-Retraining	-	-	Yes	No	Yes	No	No	No	No	No	No
EER_{ne}	4.12	5.31	6.87	5.31	6.64	4.12	5.31	5.31	4.12	5.31	5.31
EER_{wh}	23.54	22.7	21.1	24.7	20.2	24.2	23.7	18.75	16.06	**15.12**	16.24

Table 3: *Comparison of EER for different values of N_w^e between the proposed and WFM methods.*

N_w^e	0	1	2	4	6	8
WFM	22.7	11.31	9.71	7.13	6.46	6.46
$FM_{cs}(Af)$	15.12	11.66	10.82	7.49	7.12	7.07

better than DNN based FM in both FM_{MSE} and FM_{cs}. This could be due to the non-linear activation in the DNN overfitting on the training data channels and resulting in a degraded performance for the test channels. But it appears that overfitting in the case of FM_{cs} is lower than that of FM_{MSE}. It also can be observed that in FM_{cs} with just the bias transformation ($A = I$) performs better than that of FM_{cs} with DNN by $\sim$13%(relative). Relaxing the identity constraint on A improve the EER further by $\sim$7%(relative).

Table 2 also shows that the FM_{cs} with Af performs better than the baselines WFM and FM_{MSE} with Af by $\sim$50% and $\sim$25% respectively. This could be because the T-matrix used in the objective function helps the FM to focus more on the speaker dependent factors, and the T matrix retraining doesn't improve EER_{wh} significantly. Hence, the EER_{ne} for the FM_{cs} does not change with the FM.

To understand the effect of adding whisper data in enrollment, we consider the performance metrics for different values of N_w^e. Table 3 shows the comparison of EER for different values N_w^e. We compare the proposed method with the WFM because, WFM uses whispered features without FM for enrollment when $N_w^e > 0$. Hence, it acts as a lower bound on the EER for all FM methods. It is clear from the table that the EER consistently decreases when N_w^e increases. It is interesting to see that the gap in EER between the proposed method and WFM decreases as N_w^e increases.

To understand the differences in the proposed loss function and the MSE loss function, we monitor one loss when we use the other loss for the FM training. Fig. 2 shows the evolution of both the losses when (a) the MSE in MFCC space is used as the objective function for training (b) when the cosine distance in the i-vector space is used as the objective function for training. It is clear from Fig. 2(a) that the MSE loss is monotonically decreasing, but the cosine distance decreases in the beginning and

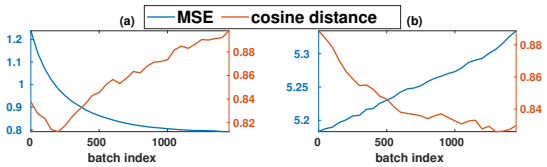

Figure 2: *Evolution of MSE and cosine distance for a batch of data during training, when the training objective is (a) MSE in MFCC space, (b) cosine distance in i-vector space.*

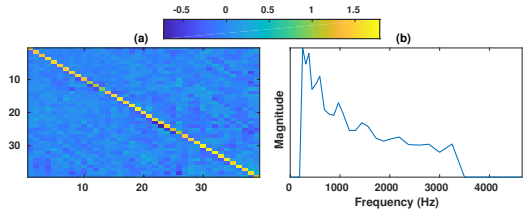

Figure 3: *(a): Learned matrix A, (b): Reconstructed spectrum from the static mfcc bias vector.*

then starts to increase. It shows that the decrease in MSE loss does not guarantee the decrease in the cosine distance. The increase in the cosine distance might indicate that the MSE based FM does not necessarily focus on increasing speaker similarity. In case of Fig. 2(b), the cosine distance decreases continuously, but the MSE increases.

Fig. 3(a) shows the trained matrix A. It is clear from the figure that the learned matrix is close to a diagonal matrix Fig. 3(b) shows the spectrum reconstructed from the bias vector using only the first 13 coefficients. It is clear from the figure that the reconstructed spectrum looks similar to the spectrum of the glottal flow. Adding this bias vector in the MFCC domain is equivalent to multiplying the glottal flow spectrum in the frequency domain. Hence, the learned bias could add the source information that is mainly missing in the whispered speech. It is also observed from Table 2 that the significant improvement in the EER happens because of the bias vector.

4. Conclusion

In this work, we proposed a novel objective function based on the cosine similarity in the i-vector space to map a whispered feature to a neutral feature, for improving the whispered SV performance. We also explore an affine transform based feature mapping. Our experiments illustrate two key advantages of the proposed objective function. Firstly, we show that learning feature mapping using the proposed objective function yields better EER compared to the baseline methods in whispered SV. We also showed that the EER for neutral SV remains unchanged, unlike baseline methods. Secondly, we showed that the affine transform with the proposed objective function performs the better than other feature mappings. We also showed that the bias of the learned affine transform is similar to the glottal flow information, typically absent in the whispered speech. Our future work includes jointly learning the T-matrix and the feature mapping for whispered SV and exploring the same objective function with other features [18, 20].

5. Acknowledgement

We thank the Department of Science & Technology, Government of India and the Pratiksha Trust for their support.

6. References

[1] D. A. Reynolds, "An overview of automatic speaker recognition technology," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, 2002, pp. 4072–4075.

[2] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[3] C. Zhang and J. H. Hansen, "Analysis and classification of speech mode: whispered through shouted," in *Eighth Annual Conference of the International Speech Communication Association*, 2007, pp. 2289–2292.

[4] X. Fan and J. H. Hansen, "Speaker identification within whispered speech audio streams," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 5, pp. 1408–1421, 2011.

[5] C. Zhang, G. S. Morrison, and P. Rose, "Forensic speaker recognition in chinese: a multivariate likelihood ratio discrimination on /i/and/y," in *Ninth Annual Conference of the International Speech Communication Association*, 2008, pp. 1937–1940.

[6] S. Adler, "Speech after laryngectomy," *The American Journal of Nursing*, vol. 69, no. 10, pp. 2138–2141, 1969. [Online]. Available: http://www.jstor.org/stable/3454024

[7] V. C. Tartter, "What's in a whisper?" *The Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 1678–1683, 1989.

[8] X. Fan and J. H. Hansen, "Speaker identification for whispered speech based on frequency warping and score competition," in *Ninth Annual Conference of the International Speech Communication Association*, 2008, pp. 1313–1316.

[9] S. T. Jovičić, "Formant feature differences between whispered and voiced sustained vowels," *Acta Acustica united with Acustica*, vol. 84, no. 4, pp. 739–743, 1998.

[10] M. J. Osfar, "Articulation of whispered alveolar consonants," Ph.D. dissertation, Urbana, Illinois, 2011.

[11] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[12] E. Variani, X. Lei, E. McDermott, I. Lopez-Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification." in *ICASSP*, vol. 14, 2014, pp. 4052–4056.

[13] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5329–5333.

[14] C. Zhang, K. Koishida, and J. H. L. Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1633–1644, Sep. 2018.

[15] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[16] X. Fan and J. H. Hansen, "Speaker identification for whispered speech using modified temporal patterns and MFCC's," in *Tenth Annual Conference of the International Speech Communication Association*, 2009, pp. 896–899.

[17] M. O. Sarria-Paja and T. H. Falk, "Strategies to enhance whispered speech speaker verification: A comparative analysis," *Canadian Acoustics*, vol. 43, no. 4, pp. 31–45, 2015.

[18] M. Sarria-Paja and T. H. Falk, "Fusion of auditory inspired amplitude modulation spectrum and cepstral features for whispered and normal speech speaker verification," *Computer Speech & Language*, vol. 45, pp. 437–456, 2017.

[19] ——, "Fusion of bottleneck, spectral and modulation spectral features for improved speaker verification of neutral and whispered speech," *Speech Communication*, vol. 102, pp. 78–86, 2018.

[20] A. R. Naini, A. Rao MV, and P. K. Ghosh, "Formant-gaps features for speaker verification using whispered speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 6231–6235.

[21] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[22] M. Sarria-Paja, M. Senoussaoui, D. O'Shaughnessy, and T. H. Falk, "Feature mapping, score-, and feature-level fusion for improved normal and whispered speech speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5480–5484.

[23] V. Vestman, D. Gowda, M. Sahidullah, P. Alku, and T. Kinnunen, "Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction," *Speech Communication*, vol. 99, pp. 62–79, 2018.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[25] F. Cummins, M. Grimaldi, T. Leonard, and J. Simko, "The chains corpus: Characterizing individual speakers," in *Proc of SPECOM*, vol. 6, 2006, pp. 431–435.

[26] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2011.

[27] A. A. Wrench, "A multichannel articulatory database and its application for automatic speech recognition," in *Proceedings 5 th Seminar of Speech Production*, 2000, pp. 305–308.

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

[29] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium, 1993*, 1993.

[30] G. N. Meenakshi and P. K. Ghosh, "Whispered speech to neutral speech conversion using bidirectional LSTMs," *Proc. Interspeech Hyderabad, India*, pp. 491–495, 2018.

[31] J. P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[32] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

[33] F. Chollet *et al.*, "Keras," 2015.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.