# FORMANT-GAPS FEATURES FOR SPEAKER VERIFICATION USING WHISPERED SPEECH

*Abinay Reddy Naini, Achuth Rao MV, Prasanta Kumar Ghosh*

Electrical Engineering, Indian Institute of Science (IISc), Bangalore- 560012, India

## ABSTRACT

In this work, we propose a new feature based on formants for whispered speaker verification (SV) task, where neutral data is used for enrollment and whispered recordings are used for test. Such a mismatch between enrollment and test often degrades the performance of whispered SV systems due to the difference in acoustic characteristics of whispered and neutral speech. We hypothesize that the proposed formant and formant gap ($FoG$) features are more invariant to the modes of speech in capturing speaker specific information compared to traditional baseline features for SV including mel frequency cepstral coefficients (MFCC) and auditory-inspired amplitude modulation features (AAMF). Whispered SV experiments with 714 speakers comprising 29232 neutral and 22932 whispered recordings reveal that the equal error rate (EER) using the proposed features is lower than that using the best baseline features by $\sim$3.79% (absolute). It was also observed that at least four whispered recordings during enrollment are required for the baseline features to perform at par with the proposed features. However, it was found that the best performing baseline features yield an EER for neutral SV task which is $\sim$1.88% higher than that using the proposed features.

***Index Terms***— whispered speech, speaker verification, formants

## 1. INTRODUCTION

A speaker verification (SV) system is designed to verify whether a given speech recording is from an enrolled speaker or not [1]. During enrollment, a speaker registers himself/herself with a few of his/her voice samples. In the test phase, the SV system needs to verify whether the test voice sample matches with that of any of the enrolled speakers. Unlike in speaker recognition, where the goal is to map a speech from an unknown speaker to the closest speaker from a set of enrolled speakers, SV system has to reject if a test speech is from an imposter [1].

While neutral speech is often used for SV purposes [2], there are several applications where SV using whispered speech is of interest. For example, using whispered speech, speakers often convey private information like a password for a biometric system [3, 4]. In some cases, criminals might whisper in a telephonic conversation to hide from the forensic authorities. For some people whose vocal folds are surgically removed, whispered speech is the only mode of communication [5]. These bring up a need for developing an SV system robust to whispered speech in addition to neutral speech.

There are several differences in the acoustics of neutral and whispered speech. For example, there is no voicing in whispered speech [4]. Low frequency formants in whispered speech are shifted compared to those in neutral speech [6]. It is also shown that hyper-articulation occurs while whispering to ensure intelligibility [7]. Despite these differences, previous works in the literature demonstrated that whispered speech contains substantial information about the content, and the speaker [4]. However, it remains a challenge to improve SV system performance when it is enrolled with only neutral speech but whispered speech is used during testing. This is primarily due to the absence of pitch and shifts in the first two formants [8, 6].

Extensive research has been done on the neutral speech based SV, referred to as neutral SV in this work. Front-end factor analysis (i-vector) [9] and Deep Neural Network (DNN) embedding [10, 11, 12] are considered to be the state-of-the-art neutral SV methods. Wu et al. [2] provided a review of various neutral SV. Unlike neutral SV, in whispered SV, enrollment is done using neutral and/or whispered speech and only whispered speech is used during test. There are a few attempts in the literature toward whispered SV. These methods can be broadly classified into two categories. In the first category, Gaussian Mixture Models (GMM) with features such as frequency warping, instantaneous frequencies, modified temporal patterns, feature mapping [13, 3, 14, 15], modified Linear frequency cepstral coefficients [16] are used. In the second category, i-vectors front end along with various features are used [9]. Sarria et al. [17] explored mean Hilbert envelope coefficients, weighted instantaneous frequencies based features for whispered SV. Sarria et al. also demonstrated the need to include data from both whispered and neutral speech in order to allow for an SV system to handle both neutral and whispered speech effectively during test. He also demonstrated The importance of a DNN based mapping from whispered speech features to neutral speech features [18] and showed an improvement in whispered SV performance. He also proposed auditory-inspired amplitude modulation features (AAMF) [19] for whispered SV. Further, he explored various fusion strategies of AAMF, spectral and bottleneck features [20] to improve the whispered SV. Vestman et al. [21] provided a survey of various whispered SV.

In this paper, we propose features based on formant and formant gaps (FoGs) along with the front-end i-vector for whispered SV. Formants based features have been used for a number of applications in the past including language recognition [22], accent detection [23] and emotion recognition [24]. In addition, a number of works on neutral SV in the past have also used formant based features. Nolan and Grigoras [25] showed that the distributions of first three formants within a speaker are similar and capture the speaker information. Over time long-term formant features are shown to be useful in SV [26]. Becker et al. used first three formants with front-end GMM-Universal Background Model (GMM-UBM) for SV [27]. Most of the works on SV using formant features deal with neutral SV [28, 29, 30]. Javier et al. [28] provided a review of formant based SV methods. To the best of our knowledge, no formant based features were proposed for whispered SV.

In whispering, speakers achieve articulatory targets more consistently unlike that in neutral speech [31]. This has often been attributed to the goal of preserving intelligibility in the absence of pitch [7]. It could be that the way the articulation during whispering changes compared to that in neutral speech is speaker specific

| Database | Num. of Speakers | | Recordings/speaker | |
|---|---|---|---|---|
| | Female | Male | Normal | Whisper |
| **TIMIT** | 192 | 438 | 10 | - |
| **wTIMIT** | 24 | 24 | 450 | 450 |
| **CHAINS** | 16 | 20 | 37 | 37 |

**Table 1**. Number of male/female speakers and recordings per speaker for all three databases considered in this work.

[32]. This, in turn, could reflect in the formants extracted from the acoustic signal. We also hypothesize that the shift in formant values from neutral to whisper could vary from one formant to another and considering gap between two consecutive formants could explicitly capture cues about the way a speaker changes his/her articulation. In fact, pitch (often used as identity of a speaker) has been estimated using formant gaps from whispered speech [4, 33]. In neutral speech, it is shown that considerable correlation exists between the first two formants and pitch because of source filter interaction [34]. Thus formants and FoGs could be unique to an individual speaker and invariant to the modes of speech (neutral or whispered), which, in turn, could capture the speaker's identity well.

The SV performance using proposed formants and FoGs features are examined using SV experiments comprising data from three databases consisting of a total of 714 number of speakers. We consider 29232 neutral speech recording and 22932 whispered recordings. When there is no whispered data in enrollment, we observe that the Equal Error Rate (EER) using the proposed features is lower than that using the best baseline features by ∼3.79% (absolute) in whisper SV although the baseline features perform better than the best performing proposed features by ∼1.88% (absolute) in neutral SV.

## 2. DATABASE

In this study, we have considered 3 different databases: (i) CHAINS [35] corpus contains 36 speakers, among them 28 are from the Eastern part of Ireland and the remaining 8 speakers are from the UK and the USA. The CHAINS data was recorded in six different conditions, including Synchronous, Fast, whispered speech etc. However, we have considered only neutral solo speech and whispered speech of 10 recordings from each of 36 subjects recorded at 44.1kHz. (ii) wTIMIT (whispered TIMIT) [36] is a large dataset comprising 20 Singaporean and 28 North American speakers, each speaking 450 recordings, both in neutral and whispered speech recorded at 44.1kHz. Although we have considered all recordings from 24 speakers and 10 recordings from each of the neutral and whispered speech have been used from remaining 24 speakers. (iii) TIMIT dataset [37] contains a total of 630 speakers from eight major dialect regions of the United States speaking only 10 neutral speech recordings, which are recorded at 16kHz. Among all TIMIT speakers, we have considered 562 speakers following the experimental setup of the work by Sarria et al. [19]. The duration of a speech utterance, when averaged over all three databases, is found to be ∼4.5 seconds. Details of the number of male and female speakers and recordings per speaker are given in Table 1. We resampled all the recordings to a sampling frequency ($f_s$) of 16kHz.

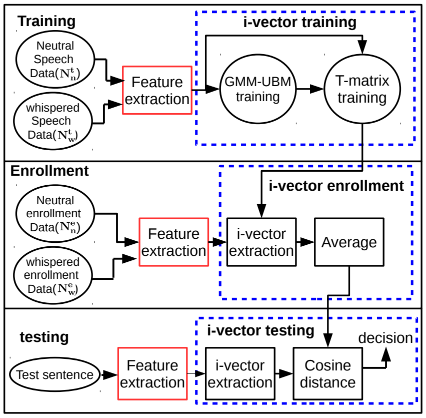

**Fig. 1**. Block diagram of speaker verification system. Feature extraction step is shown in red and i-vector step is shown in blue.

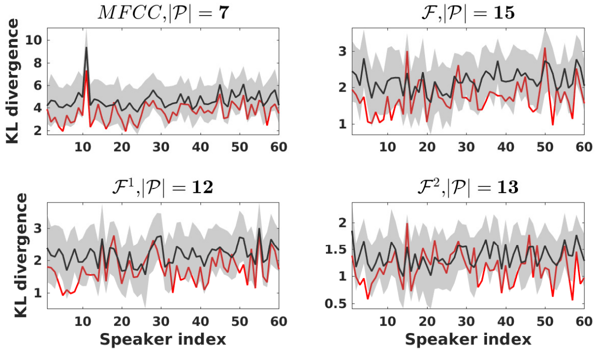

**Fig. 2**. KL divergence between neutral and whispered feature distributions both within (red) and across (black) subjects using different features, namely, MFCC, $\mathcal{F}$, $\mathcal{F}^1$, $\mathcal{F}^2$. Gray shaded region indicates 1.5 times the standard deviation interval around the mean for the across subject KL divergence.

## 3. SPEAKER VERIFICATION USING PROPOSED FEATURE

Fig. 1 shows a typical SV system. It involves three stages comprising training, enrollment, and testing. Each stage involves feature extraction step followed by the front-end i-vector step. The number of whispered and neutral utterances used in the training step is denoted by $N_w^t$ and $N_n^t$. The number of whispered and neutral utterances per speaker used in enrollment step is denoted by $N_w^e$ and $N_n^e$. The proposed feature extraction step and the i-vector step are explained in detail below.

### 3.1. Proposed Formant Gap (FoG) features

To compute the proposed features, we divided the speech signal into frames of window length $N_w$ with a shift of $N_s$. For each window, we computed five formants indicated by a vector of $\mathcal{F} = [f_1, f_2, f_3, f_4, f_5]$, where $f_i$ indicates the $i$-th formant. In this work, we explored first ($f_i^1$) and second order ($f_i^2$) formant gaps defined below:

$$f_i^1 = f_{i+1} - f_i, \; 1 \leq i \leq 4 \tag{1}$$

$$f_i^2 = f_{i+1}^1 - f_i^1, \; 1 \leq i \leq 3 \tag{2}$$

Let $\mathcal{F}^1 = \{f_i^1; 1 \leq i \leq 4\}$, $\mathcal{F}^2 = \{f_i^2; 1 \leq i \leq 3\}$. Two types of feature vectors are constructed using FoGs, namely, $FoG_1 = [F, \mathcal{F}^1]$ and $FoG_2 = [\mathcal{F}, \mathcal{F}^1, \mathcal{F}^2]$.

Features that capture a speaker specific information and are invariant to the modes (whispered, neutral) of speech, would be appropriate for whispered SV, as the modes of speech in enrollment and test are different in whispered SV. In order to understand the difference between the distribution of the proposed features in whispered and neutral speech, we conducted an illustrative experiment, in which, we trained a speaker specific GMM for whispered and neutral speech features separately. To compare the differences between the distribution within and across speakers, we computed two kinds of symmetric Kullback-Leibler (KL) divergence [38]. First, to measure the difference between the whispered and neutral feature distribution within a speaker, we computed the KL divergence between $i$-th speaker's neutral GMM ($N_i$) and whispered GMM ($W_i$) denoted by $D(N_i|W_i)$. To measure the difference between the neutral GMM and the whispered GMM across speakers, we computed the average and standard deviation of KL divergence between the $N_i$ and $W_{j\neq i}$ as follows:

$$M_{KL}(i) = \frac{1}{N-1}\sum_{j} D(N_i|W_{j\neq i}), \tag{3}$$

$$\sigma_{KL}(i) = \sqrt{\frac{1}{N-1}\sum_{j}(D(N_i|W_{j\neq i}) - M_{KL}(i))^2} \tag{4}$$

where $1 \leq i \leq N$ and $N$ is the number of speakers. We find the set of speakers $\mathcal{P}$ as follows: $\mathcal{P} = \{i : D(N_i|W_i) < M_{KL}(i) - 1.5 \times \sigma_{KL}(i)\}$. The cardinality of $\mathcal{P}$ ($|\mathcal{P}|$) indicates the number of speakers whose feature distribution in neutral speech is much closer to that of their corresponding whispered speech compared to other speakers' whispered speech. For this illustrative experiment, we have considered 60 speakers comprising 24 from WTIMIT and 36 from CHAINS corpora. The GMMs are trained with 10 sentences from neutral and whispered speech for each speaker. Fig. 2 shows the plot of $M_{KL}(i)$ with gray shaded area indicating 1.5 $\sigma_{KL}$ interval and $D(N_i|W_i)$ for the proposed FoG features and MFCC. The $|\mathcal{P}|$ for different features is mentioned on top of individual plots in Fig. 2. Higher value of $|\mathcal{P}|$ would imply greater separation between a speaker and every other speakers in the proposed feature space. It is clear from Fig. 2 that formant based features achieve higher value of $|\mathcal{P}|$ compared to that from MFCC suggesting that the proposed formant and FoGs could capture speaker information irrespective of the modes (whispered and neutral) of speech. It is interesting to observe that $|\mathcal{P}|$ using $\mathcal{F}^2$ is not significantly different from that using $\mathcal{F}^1$. It could be that $\mathcal{F}^2$ does not carry speaker specific information complementary to that using $\mathcal{F}^1$.

### 3.2. Front-end i-vector step

The i-vector extraction is a dimensionality reduction procedure using factor analysis [9]. In the i-vector training step, we considered extracted features of dimension $F$ to train a GMM-UBM with $C$ mixtures using the Expectation-Maximization algorithm [39], and by concatenating all mixture means, a speaker and channel independent super vector ($m_0$) of dimension $CF$ is obtained. Then each speaker or channel dependent super vector ($M_s$) with dimension $CF$ is modeled as $M_s = m_0 + T \cdot w$. Here we train a tall and low rank $T$ (Total variability) matrix with dimension $CF \times d$ by assuming that i-vector ($w$) follow standard normal distribution, as explained in [40]. The $T$ matrix is used to map a low dimensional i-vector of dimension $d$ to a high dimensional $M_s$. Using any speaker dependent super vector $M_s$, $m_0$ and $T$ matrix we can compute i-vector of the speaker, as explained in [9]. In the enrollment step, we extracted i-vectors for each neutral and whisper speech recordings in a similar way, and by taking an average of all these i-vectors of a speaker we obtained one final i-vector for each speaker. During test, we first computed i-vector for test recording, then using linear discriminant analysis (LDA) we reduced the dimension of i-vectors, following which we computed cosine kernel distance between test and enrolled i-vectors for making an SV decision.

| | Num. of Speakers/database | | | Total Recordings | |
|---|---|---|---|---|---|
| | TIMIT | wTIMIT | CHAINS | Ne. | Wh. |
| UBM training | 462 | 0 | 0 | 3696 | 0 |
| T matrix training | 462 | 24 | 0 | 9996 | 10800 |
| LDA training | 462 | 24 | 0 | 9996 | 10800 |
| Enrollment | 100 | 24 | 36 | 1280 | 480 |
| Testing | 0 | 24 | 36 | 120 | 120 |

**Table 2**. Number of speakers and the total number of recordings per database for training, enrollment, and testing.

## 4. EXPERIMENTAL SETUP

In the experimental stage, we divided recordings from all three datasets into training and testing speakers. 462 speakers from the TIMIT database which contains only neutral speech and 24 speakers (both whispered and neutral speech) from the wTIMIT database are selected for training. Details of train/test split of speakers and number of recordings per speaker are provided in Table 2. In the enrollment phase, we fixed eight neutral utterances per test speaker and for the number of whispered utterances, we have considered two cases. In the first case, we assume that no whispered utterances for the test speakers are available ($N_w^e = 0$). In the second case, we varied the number of whispered utterances used $N_w^e \in \{2, 4, 6, 8\}$. In the test phase, we have two scenarios. In the first scenario, we tested with two whispered utterances from each test speaker in both the above cases. In the second scenario, we tested with two neutral utterances from each speaker only in the first case ($N_w^e = 0$). The sentences used in the test phase do not overlap with those used in the enrollment phase. Below we describe the features and the evaluation metric used in this work.

### 4.1. Features

In this sub-section, we explain the implementation details of the proposed features. We also provide a detailed explanation of the three baseline features considered. Our experiments use these features along with the front end i-vector step for SV.

### 4.2.1 Proposed FoG features

We divided the speech signal into frames using a window duration of $N_w(= 25ms)$ and shift of $N_s(= 10ms)$. For each window, five formants are extracted using an algorithm based on peak picking on differential phase spectrum, proposed by Baris Bozkurt et al. [41]. Unlike the other algorithms, the algorithm in [41] has been shown to estimate formants, in particular $4^{th}$ formant, with high precision. The formants are normalized by $f_s/4$. The FoG features from formants are extracted using eq. (1) and (2).

### 4.2.2 Baseline features

**MFCC:** MFCCs are widely used features in different speech applications. Speech signal, pre-emphasized with a filter coefficient of 0.97 is used to obtain a 13-dimensional MFCC feature vector. Features were computed over $25ms$ window with a shift of $10ms$. To add temporal dynamics to the feature vector, velocity and acceleration coefficients were computed resulting in a 39-dimensional feature vector [19].

**Auditory-inspired amplitude modulation features (AAMF):** AAMF features [19] assumed that the speech frame is a result of multiplying a low frequency modulating signal by the high-frequency carrier. Hence, the modulation spectrum encodes the rate of change of long-term speech temporal envelopes. First speech frames were transformed into magnitude squared of short-time Fourier transform and frequency components were grouped to get 27 subbands according to the perceptual Mel-scale [42]. Each subband time series was transformed into magnitude squared of short-time Fourier transform and modulation frequency bins were further grouped into eight subbands using logarithmically-spaced triangular bandpass filters distributed over a range of 0.01-80 Hz modulation frequency. This resulted in a feature dimension of $27 \times 8$ (216). Finally, logarithm of these features were computed. The feature dimension was reduced to 40 using principal component analysis [43].

**Deep neural network(DNN) based feature mapping:** We considered both MFCC and AAMF features of all training speakers whose whispered and neutral speech for the same utterance was available. We have used Dynamic time warping (DTW) to align whispered and neutral speech. Following this, two DNNs were trained to perform whispered to neutral speech feature mapping for MFCC and AAMF [18]. In the testing phase, whispered features were transformed using the DNNs before computing i-vectors. These transformed features are denoted by $MFCC_{DNN}$ and $AAMF_{DNN}$ respectively.

### 4.2. Evaluation metrics

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 enrolled speakers are equal [44].

## 5. RESULTS AND DISCUSSION

Initially, we explored the importance of the different combinations of proposed features. First three rows in Table 3 show a comparison of EER using three combinations of proposed features in zero whispered enrollment condition. It is clear from the table that combination of $\mathcal{F}$ and $\mathcal{F}_1$ features $(FoG_1)$ performs the best in the case of whispered test condition and $\mathcal{F}$ performs the best in neutral test conditions among proposed features. The EER using $(FoG_1)$ features is lower than that using formant $(\mathcal{F})$ features in whispered test condition. This could be due to a shift in the first two formants in whispered speech compared to the neutral speech, which is better captured in formant gaps.

In the neutral test condition, the formant features achieve the least EER compared to other proposed features. This could be due to its capacity in separating speakers in the feature space which is reflected in the observations made from Fig.2. We observe that the EER for $(FoG_2)$ is greater than $(FoG_1)$, in both neural and whispered test condition. This could be because $\mathcal{F}^2$ may not carry speaker information complementary to those carried by $\mathcal{F}$ and $\mathcal{F}^2$ similar to the observations made in Figure. 2. It is also interesting to observe that $FoG1$ increases the EER by $1.52$ in the neutral SV case compared to $\mathcal{F}$ unlike a drop of $9.42$ in EER for whispered SV. This could be because $\mathcal{F}$ features capture speaker specific information better in neutral speech while $FoG1$ features capture speaker information better irrespective of the modes (whisper and neutral) of speech.

Table 3also show EER using baseline features for zero whisper enrollment data in order to compare with those using the proposed features. It is clear from the table that the feature mapping on the baseline feature $(MFCC_{DNN}$ and $AAMF_{DNN})$ reduces the EER compared to original counterparts (MFCC and AAMF) for the whisper test condition. It is also clear from the table that the proposed feature performs better than all the baseline features in whisper test condition. In particular, five dimensional $\mathcal{F}$ features yields an EER similar to that using 39-dimensional MFCC features in both whisper and neutral test condition suggesting that the $\mathcal{F}$ features compactly represent speaker specific information compared to MFCC. This could is due to its better invariant nature from whispered to neutral speech in the feature domain. The better performance of AAMF in neutral test case is due to its better separation of speakers across the different channels. One of the reasons for its poor performance in whispered test condition could be due to its sensitivity towards speaking rate, which varies significantly from whispered to neutral speech.

| | | Test condition | |
|---|---|---|---|
| features | | whisper | Neutral |
| proposed | $\mathcal{F}$ (5) | 22.42 | 6.28 |
| | $FoG_1$ (9) | **13.00** | 7.8 |
| | $FoG_2$ (12) | 14.98 | 9.14 |
| baseline | MFCC (39) | 22.47 | 6.25 |
| | AAMF (40) | 19.81 | **4.4** |
| | MFCC$_{DNN}$ (39) | 17.01 | - |
| | AAMF$_{DNN}$ (40) | 16.79 | - |

**Table 3**. Comparison of EER using the proposed features with baseline features for both whispered and neutral test conditions with $N_w^e = 0$. Numbers in bracket indicate the dimension of the feature vector.

| $N_w^e$ | 0 | 2 | 4 | 6 | 8 |
|---|---|---|---|---|---|
| AAMF$_{DNN}$ | 17.01 | 14.14 | **8.61** | **6.14** | **4.78** |
| $FoG_1$ | **13.00** | **10.82** | 9.68 | 8.88 | 8.66 |

**Table 4**. Comparison of EER using the proposed features with that using best baseline features for different values of $N_w^e$ in whisper test condition.

Table 4 shows a comparison between the best baseline and the $FoG_1$ when the $N_w^e$ is varied in whisper test condition. It is clear from the table that the the SV using baseline features requires at least four whisper recordings in the enrollment phase for it to perform better than the proposed features. This, in turn, suggests that the proposed features are robust to the modes (whisper and neutral) of speech for SV applications.

## 6. CONCLUSION

In this paper, we proposed formants and formant-gaps (FoGs) feature for whispered speaker verification. Experiments show that among the proposed features, formants with 1st order formant gaps perform better than other formant based features. We showed that the proposed features perform 3.79% (absolute) better than the best baseline in absence of whisper data in the enrollment stage. We also observed that the proposed $FoG$ feature performs better than the best baseline feature, namely AAMF, when two or less whispered recordings are available for enrollment in whisper speech test condition. However, the performance of $FoG$ features is slightly lower than AAMF features in neutral test conditions. Further investigation is required on large corpus, to improve the neutral SV performance using proposed FoG features, also to explore different feature mapping methods for FoG features.

**Acknowledgement**: We thank the Pratiksha Trust for their support.

## 7. REFERENCES

[1] Bishnu 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.

[2] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li, "Spoofing and countermeasures for speaker verification: A survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[3] Xing Fan and John HL Hansen, "Speaker identification within whispered speech audio streams," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 5, pp. 1408–1421, 2011.

[4] Vivien C Tartter, "Whats in a whisper?," *The Journal of the Acoustical Society of America*, vol. 86, no. 5, pp. 1678–1683, 1989.

[5] Sol Adler, "Speech after laryngectomy," *The American Journal of Nursing*, vol. 69, no. 10, pp. 2138–2141, 1969.

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

[7] Megan J Osfar, *Articulation of whispered alveolar consonants*, Ph.D. thesis, Urbana, Illinois, 2011.

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

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

[10] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez-Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification.," in *ICASSP*, 2014, vol. 14, pp. 4052–4056.

[11] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, Oct 2015.

[12] 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.

[13] Xing Fan and John HL 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.

[14] Xing Fan and John HL Hansen, "Speaker identification for whispered speech using modified temporal patterns and mfccs," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[15] Xing Fan and John HL Hansen, "Acoustic analysis and feature transformation from neutral to whisper for speaker identification within whispered speech audio streams," *Speech communication*, vol. 55, no. 1, pp. 119–134, 2013.

[16] Xing Fan and John HL Hansen, "Speaker identification with whispered speech based on modified lfcc parameters and feature mapping," in *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2009, pp. 4553–4556.

[17] Milton O Sarria-Paja and Tiago 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, 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.

[19] Milton Sarria-Paja and Tiago 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.

[20] Milton Sarria-Paja and Tiago H Falk, "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.

[21] Ville Vestman, Dhananjaya Gowda, Md Sahidullah, Paavo Alku, and Tomi 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.

[22] Javier Franco-Pedroso and Joaquin Gonzalez-Rodriguez, "Linguistically-constrained formant-based i-vectors for automatic speaker recognition," *Speech Communication*, vol. 76, pp. 61–81, 2016.

[23] Liu Wai Kat and Pascale Fung, "Fast accent identification and accented speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999. Proceedings (ICASSP)*. IEEE, 1999, vol. 1, pp. 221–224.

[24] Yongjin Wang and Ling Guan, "Recognizing human emotional state from audiovisual signals," *IEEE transactions on multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[25] Francis Nolan and Catalin Grigoras, "A case for formant analysis in forensic speaker identification," *International Journal of Speech Language and the Law*, vol. 12, no. 2, pp. 143, 2005.

[26] Michael Jessen and Timo Becker, "Long-term formant distribution as a forensic-phonetic feature.," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2378–2378, 2010.

[27] Timo Becker, Michael Jessen, and Catalin Grigoras, "Forensic speaker verification using formant features and gaussian mixture models," in *Ninth Annual Conference of the International Speech Communication Association*, pp. 1505–1508.

[28] Javier Franco-Pedroso and Joaquin Gonzalez-Rodriguez, "Linguistically-constrained formant-based i-vectors for automatic speaker recognition," *Speech Communication*, vol. 76, pp. 61 – 81, 2016.

[29] Kirsty McDougall, "Dynamic features of speech and the characterization of speakers," *Int. J. Speech Lang. Law*, vol. 13, pp. 89–126, 2006.

[30] Cuiling Zhang, Geoffrey Stewart Morrison, and Philip 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*, pp. 1937–1940.

[31] Man Gao, *Tones in whispered Chinese: articulatory features and perceptual cues*, Ph.D. thesis, University of Victoria, 2002.

[32] Aravind Illa, Prasanta Kumar Ghosh, et al., "A comparative study of acoustic-to-articulatory inversion for neutral and whispered speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5075–5079.

[33] Ian V Mcloughlin, Hamid Reza Sharifzadeh, Su Lim Tan, Jingjie Li, and Yan Song, "Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 6, no. 4, pp. 12, 2015.

[34] Vinay Kumar Mittal, B Yegnanarayana, and Peri Bhaskararao, "Study of the effects of vocal tract constriction on glottal vibration," *The Journal of the Acoustical Society of America*, vol. 136, no. 4, pp. 1932–1941, 2014.

[35] Fred Cummins, Marco Grimaldi, Thomas Leonard, and Juraj Simko, "The chains corpus: Characterizing individual speakers," in *Proc of SPECOM*, 2006, vol. 6, pp. 431–435.

[36] Boon Pang Lim, *Computational differences between whispered and non-whispered speech*, Ph.D. thesis, University of Illinois at Urbana-Champaign, 2011.

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

[38] Solomon Kullback and Richard A Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[39] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[40] Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE transactions on speech and audio processing*, vol. 13, no. 3, pp. 345–354, 2005.

[41] Baris Bozkurt, Thierry Dutoit, Boris Doval, and Christophe d'Alessandro, "Improved differential phase spectrum processing for formant tracking," in *Eighth International Conference on Spoken Language Processing*, 2004.

[42] Juanjuan Xiang, David Poeppel, and Jonathan Z Simon, "Physiological evidence for auditory modulation filterbanks: Cortical responses to concurrent modulations," *The Journal of the Acoustical Society of America*, vol. 133, no. 1, pp. EL7–EL12, 2013.

[43] Michael E Tipping and Christopher M Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

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