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

Competency Evaluation in Voice Mimicking Using Acoustic Cues

Abhijith G., Adharsh S., Akshay P. L., Rajeev Rajan

College of Engineering, Trivandrum,
Thiruvananthapuram, Kerala, India

rajeev@cet.ac.in

Abstract

The fusion of i-vector with prosodic features is used to identify the most competent voice imitator through a deep neural network framework (DNN) in this paper. This experiment is conducted by analyzing the spectral and prosodic characteristics during voice imitation. Spectral features include mel-frequency cepstral features (MFCC) and modified group delay features (MODGDF). Prosodic features, computed by the Legendre polynomial approximation, are used as complementary information to the i-vector model. Proposed system evaluates the competence of artists in voice mimicking and ranks them according to the scores from a classifier based on mean opinion score (MOS). If the artist with the highest MOS is identified as rank-1 by the proposed system, a hit occurs. DNN-based classifier makes the decision based on the probability value on the nodes at the output layer. The performance is evaluated using top X-hit criteria on a mimicry dataset. Top-2 hit rate of 81.81% is obtained for fusion experiment. The experiments demonstrate the potential of i-vector framework and its fusion in competency evaluation of voice mimicking.

Index Terms: meter, poem, fusion, timbral, i-vector -0.1cm

1. Introduction

A mimic is a performer who imitates the voice and mannerisms of others. Imitator modifies the position of articulators like lips, tongue, speech duration, and pitch to reproduce another person's voice [1] during the mimicry. A competent imitator controls articulators to get a closer voice to that of the target. F1-F2 plot of few speech frames from target speaker and two imitators are shown in Figure 1. From Figure 1, it is worth mentioning that formant locations of the target and best imitator are closer than that of the worst imitator [2]. The proposed system evaluates the competency (the ability to mimic the voice with maximum similarity) of imitators in voice mimicking by analyzing the mimicked speech using acoustic approach rather than the mannerism.

Professional voice imitator can approximate the system parameters of a well-known target speaker [3]. A whole text unit and one selected word which occurs in all the recordings have been analyzed using auditory, acoustic and phonetic features during imitation [3]. In [4], the authors point out that the inclusion of prosodic features for an automatic speaker recognition system requires careful consideration of the risk of impersonation. Gomati *et al.* [5], analyse the imitated speech using various features at segmental, suprasegmental and subsegmental levels for both good and poor imitation cases. Words and phrases related to the target speaker make it easier for the audience to identify the imitated voice [6]. The mean and dynamics of pitch, vocal tract acoustic characteristics and the glottal source characteristics are very much important during voice imitation [7]. It is worth mentioning that in most of the experiments, effectiveness of voice mimicking is evaluated using the same utterance as that of the target. But the ideal system should be able to find the best impersonator, even if the text uttered is different.

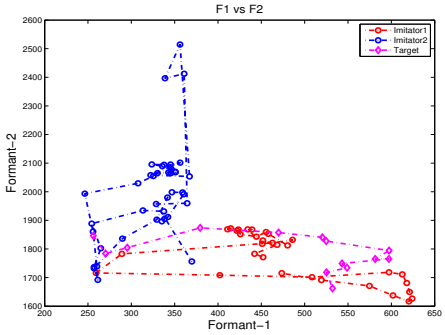

Figure 1: *F1-F2 plot for a word 'ithupōle'*

Conventionally, the spectrum-related features used in recognition take into account merely the magnitude information. However, there is often additional information concealed in the phase, which could be beneficial for recognition [8, 9, 10]. The effectiveness of group delay features in speech/speaker recognition [11, 12] and the results in [9, 10] motivated us to use group delay functions in the proposed task. More recently factor analysis techniques have been considered as a front-end to form a low-dimensional total-variability subspace that can be classified more efficiently than conventionally used high-dimensional super vectors [13, 14]. The proposed framework is a novel attempt to apply the i-vector framework in competency evaluation of voice mimicking. The artists who imitate the voice of celebrity (target) are referred to as imitators and all other terminologies are followed from [15] .

The organization of rest of the paper is as follows. Section 2 gives an overview of the proposed system. Performance evaluation is described in Section 3, followed by the analysis of results in Section 4. The paper is concluded in Section 5.

2. Overview of the proposed system

Block diagram of the proposed system is shown in Figure 2. MFCC, MODGDF, i-vectors and prosodic features are computed in the front-end. In the first phase, DNN is trained using i-vectors computed from MFCCs of the celebrity/target. Each node in the output of DNN corresponds to one celebrity. During the testing phase, the probability value (score) at the corresponding target node in the output layer is examined after inputting all the mimicked versions. The artist, who is getting the highest probability score at the target node, is assigned first rank and grades the remaining artists relatively. For instance, assume, we want to find the best voice mimicking artist (among five) in mimicking first celebrity (target). The extracted features from five artists are fed to the trained classifier, one after the other and examine the score obtained at the first node of the

http://dx.doi.org/10.21437/Interspeech.2020-1790

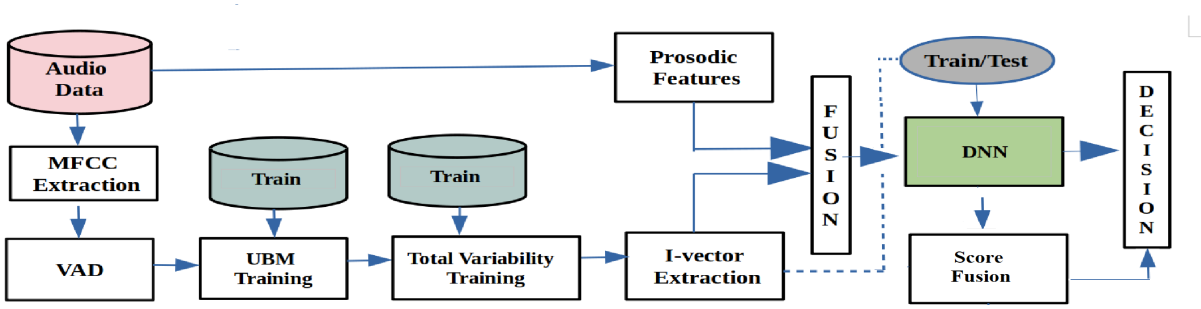

Figure 2: *Block diagram of the the proposed system*

output layer. The artist, who gets the maximum value at the node, is taken as the best artist who mimicked the first celebrity. Experiment is extended to i-vectors computed from MODGDF, feature-level and score-level fusion of both, and finally the fusion of prosodic features with i-vectors, in successive phases.

2.1. Feature Extraction

2.1.1. MFCC

MFCCs are employed in numerous perceptually motivated audio classification tasks due to the capacity to capture "global" spectral envelope properties. 20-dimensional MFCC are frame-wise computed using frame-length of 30 ms and frame-shift of 10 ms.

2.1.2. Modified group delay functions (MODGDF)

Earlier studies have already established the significance of short-term phase spectrum in speech and music processing applications [11, 16, 17]. The group delay function $\tau(e^{j\omega})$ for a discrete time signal $x[n]$ is defined by:

$$\tau(e^{j\omega}) = -\frac{d\{\arg(X(e^{j\omega}))\}}{d\omega}. \quad (1)$$

where ω is the angular frequency, $X(e^{j\omega})$ is the Fourier Transform (FT) of the signal $x[n]$ and $\arg(X(e^{j\omega}))$ is the phase function. The group delay function of minimum phase signals can be computed directly from the signal by [18].

$$\tau(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|X(e^{j\omega})|^2} \quad (2)$$

where the subscripts R and I denote the real and imaginary parts, respectively. $Y(e^{j\omega})$ represents Fourier transform n$x[n]$. $|X(e^{j\omega})|^2$ in the denominator, makes the group delay function noisy for non-minimum phase signals [19]. The denominator is replaced by its spectral envelope, $S(e^{j\omega})$ to mask the spiky nature. Modified group delay function (MODGD) $\tau_m(e^{j\omega})$ of a minimum phase signal $x[n]$ is obtained as

$$\tau_m(e^{j\omega}) = (\frac{\tau_c(e^{j\omega})}{|\tau_c(e^{j\omega})|})(|\tau_c(e^{j\omega})|)^{\alpha}, \quad (3)$$

where,

$$\tau_c(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|S(e^{j\omega})|^{2\gamma}}. \quad (4)$$

α and γ are introduced to control the dynamic range off MODGD. Group delay function and MODGD functions of a speech frame in Figure 3(a) are shown in Figure 3(b) and Figure 3(c), respectively. MODGD is converted to cepstral features (MODGDF) by decorrelating with DCT [20].

2.1.3. I-Vectors

The method of modelling the Gaussian mixture model (GMM) super vectors has achieved superior speaker recognition performance in recent works. In i-vector system [21], the high dimensional GMM super vector space (generated from concatenating all the mean values of GMM) is mapped to low dimensional space called total variability space. The target utterance GMM is adapted from a universal background model (UBM) using the eigenvoice adaption introduced in [22]. The target GMM super vector can be viewed as shifted from the UBM. Formally, a target GMM super vector M can be written as:

$$M = m + Tw \quad (5)$$

where m represents the UBM super vector, T is a low dimensional rectangular total variability (TV) matrix, and w is termed as i-vector. Using training data, the UBM and TV matrix is modeled by expectation maximization. In the E-step, w is considered as a latent variable with normal prior distribution $N(0, I)$. Eventually, the i-vectors will be estimated as the mean of posterior distribution of w, that is [21],

$$w(u) = (I + T^T\Sigma^1.N(u).T)^1T^T\Sigma^1S(u) \quad (6)$$

where for utterance u, the terms $N(u)$ and $S(u)$ represent zeroth and centralized first order Baum-Welch statistics respectively, and Σ is the covariance matrix of UBM. 10 dimensional i-vectors of MFCC (i_{MFCC}) and MODGDF (i_{MODGDF}) have been computed for each utterance for the proposed experiment.

2.1.4. Prosodic Features (P)

In the proposed work, pitch and energy features are extracted using Legendre polynomial basis. Continuous prosodic contour modelling based on Legendre polynomial expansions has been successfully used in the field of language identification [23] and quantitative phonetics [24]. Pitch and energy contours computed with 10 ms intervals are broken into segments (syllable-like regions) and approximate each segment by Legendre polynomial expansions. An approximation of pitch and energy contour is performed for each segment by taking the M (5 for the current experiment) leading terms in a Legendre polynomial expansion. Each contour $f(t)$ (where t represents time) is approximated by [25]

$$f(t) = \sum_{1=0}^{M} a_i P_i(t) \quad (7)$$

where $P_i(t)$ is the i^{th} Legendre polynomial. We compute six coefficients (a_i) to represent pitch contour, six coefficients to represent energy contour and one term for the segment duration and thus making a 13-dimensional feature vector for each

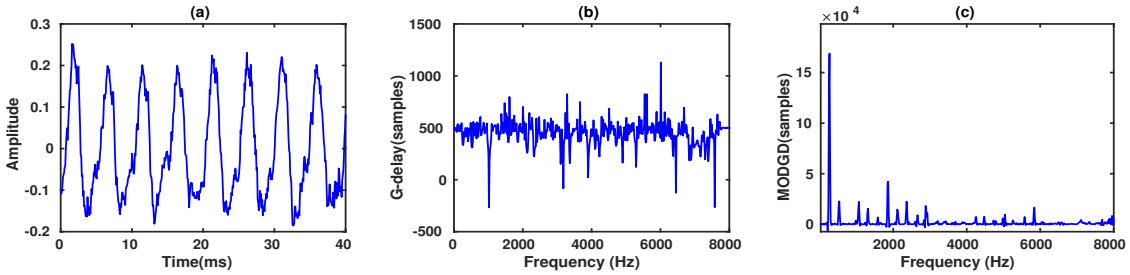

Figure 3: *(a) Frame of speech, (b) group delay function computed for (a), (b) Modified group delay function computed the frame in (a).*

segment. Average taken over all the segments is appended to the i-vector for the fusion scheme. A python framework DisVoice [26] is employed to compute the prosodic feature set (P).

2.2. Deep neural network

Our DNN is based on 4 hidden layers with number of nodes 64, 64, 64, and 32 per layer. Stochastic gradient descent (SGD) algorithm is used for the optimization. Relu and softmax have been chosen as the activation function for hidden layer and output layer, respectively. Training is carried out using target data for 6000 epochs with batch size of 512 and learning rate of 0.01. 5% of data is used for validation.

3. Performance evaluation

Proposed system is evaluated on the dataset used in the study of Mary *et al.* [1]. 22 celebrities were tested against professional artists for during voice mimicking. Artists were given the flexibility to speak any text of their choice while mimicking the celebrities in the Malayalam Language. Celebrity speeches were collected from entertainment programmes and musical shows, whereas the mimicked versions were recorded in a studio environment [1]. Our study is focused on the text-independent part, attempting to study the role of voicing characteristics in mimicry. Artist-1 to Artist-5 are the assigned labels for imitators. It does not mean that the person with label, Artist-1 is same for all the mimicking versions. Some artists may not be able to imitate certain targets effectively. Since all the artists are not comfortable to imitate all the target voice, 26 artists have participated in the proposed experiment.

3.1. Perceptual Evaluation

A perception test is conducted using twenty listeners for identifying the best imitator who mimics the celebrity well. All the listeners are presented with utterances of the target (celebrity) voice and mimicking test utterance from the imitator (artist). Listeners are asked to grade the performance of each artist in mimicking the celebrities by choosing one among the five opinion grades varying from highly dissimilar to highly similar. These grades are later converted to a numerical score as highly dissimilar (1), dissimilar (2), somewhat similar (3), similar (4), highly similar (5). Mean opinion score for each mimicked utterance is computed by taking the average of the scores given by all the twenty listeners. The artist who gets the maximum score is identified as the best one in mimicking the celebrity well.

3.2. Experimental Setup

I-vectors are computed using 128 mixture GMM from MFCC and MODGDF using Alize tool-kit [27]. UBM-GMM model is trained using features computed from the auxiliary database comprising audio samples other than the celebrities in the corpus. TV matrix is also estimated during the successive phase using the training data from targets. The experiment is carried out in five phases, namely, i_{MFCC}, i_{MODGDF}, i_{MFCC} +i_{MODGDF} (feature fusion), i_{MFCC} • i_{MODGDF} (score fusion), and finally i_{MFCC} +i_{MODGDF} + P (feature fusion). The final score S_f, in the score fusion is obtained by,

$$S_f = \beta S_{mf} + (1 - \beta) S_{mdf} \tag{8}$$

where S_{mf}, S_{mdf}, β represent i_{MFCC} score, i_{MODGDF} score and weighting constant, respectively. β = 0.7 is empirically chosen in our experiment.

Based on the probability score, computed from DNN, the system ranks artists from the best mimic to worst (Rank 1 to Rank 5). If the artist who is ranked first by the proposed system, matches with the outcome of MOS, a hit occurs and assumes that the system correctly identified the best mimicking speaker. The performance is evaluated using the top-X hit rate criteria on five artists who imitate 22 celebrities. The top-X hit rate reports the proportion of queries for which $r_i \leq X$, where r_i denotes the rank of an imitator given by the proposed system [28]. One point is scored for a hit in the top X outcome and zero is scored otherwise. For instance, if the proposed system rates the best imitator (as per MOS test) with rank-2, it is counted in the results of the top-2 hit rate.

4. Results and Analysis

The impersonator changes the shape of his vocal tract to adjust the frequency and bandwidth of the spectral peaks and dip while imitating a voice [7]. In addition to the correlation between the formant frequencies of target and imitator, it is apparent that there is a strong tendency of the imitator to adjust pitch trajectory during imitation. Prosodic features, computed from the segmented intonational phrases, are employed in evaluating the quality of mimicked speech in [1]. It is seen that scores are very close with a slight mileage for the best performing artist. A possible cause for this is the non-discriminatory nature of the selected features. However, both spectral and prosodic features play a crucial role in performance evaluation during voice mimicking, which motivated us to apply fusion scheme in our experiment.

Mean opinion score and scores obtained for the fusion system i_{MFCC} +i_{MODGDF} + P for entire targets are shown in Figure 4(a) and Figure 4(b), respectively. One can easily identify MOS obtained for each artist in mimicking one celebrity during the perception test from the Figure 4(a). By comparing Figure 4(a) and 4(b), the number of hit with high MOS can be identified. It is noticed that out of 22 cases, 11 are identified correctly. The efficacy of i-vector and its fusion is illustrated in Figure 5, with scores of various schemes and MOS for a celebrity. It can be

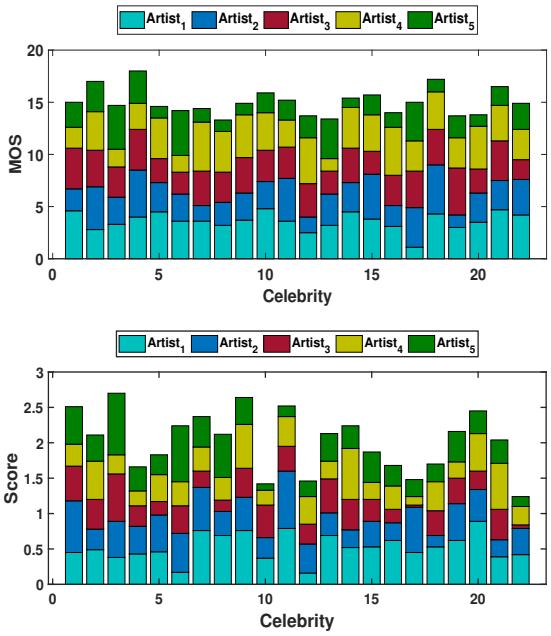

Figure 4: *(a) MOS obtained for all artists in mimicking 22 celebrities (b) Scores obtained for i_{MFCC} + i_{MODGDF} + P (Feat Fusion) for all artists in mimicking 22 celebrities.*

seen that the feature fusion of i-vectors and prosody provides maximum score for the best artist identified by the MOS.

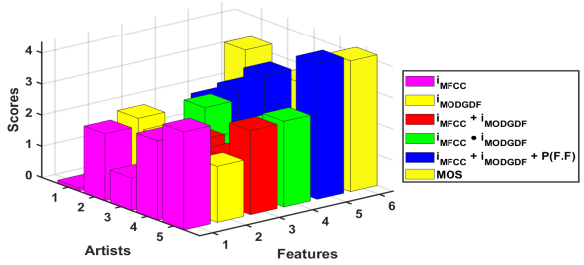

Figure 5: *Scores (Scaled) with MOS for a Celebrity.*

Overall results are tabulated in Table 1. The identification accuracy with the Rank-1 margin for the i_{MFCC} and i_{MODGDF} are 40.90% and 31.81%, respectively. The top-2 hit rate improved to 68.18% and 63.63%, for i_{MFCC}, i_{MODGDF} respectively. By examining the scores of the entire experiment, we could find that there are test cases where i_{MFCC} detects best artist correctly when i_{MODGDF} fails and vice-versa. The feature level resulted in an improvement with top-1 hit rate with accuracy of 45.45%. Even though there is no improvement for score level fusion in top-1 as compared to i_{MFCC}, top-2 rate increased. During the prosodic fusion, the performance of i_{MFCC} +i_{MODGDF}+ P improved significantly with top-2 hit rate of 81.81%. It is reasonable to say that the prosodic features acted as complementary information in decision making.

It is well established that higher-order formants are important in discriminating between speakers [29]. The ability of the group delay features to emphasize higher-order formants is investigated earlier. Modified group delay functions computed for one frame of the consonant '[p] 'for both target and best imitator is shown in Figure 6. The peaks in the MODGD plot corresponds to formant locations. It can be seen that formant locations are more or less the same for target and best imitator. Figure 7 visualizes the individual output vectors produced by the snippets from 7 celebrities for the hidden layer of the network using t-SNE [30]. The plot corresponds to i_{MFCC} +i_{MODGDF}+ P. Note that there is good clustering of target identities (as represented with colour) and a general separation of targets for i-vector-prosody feature space.

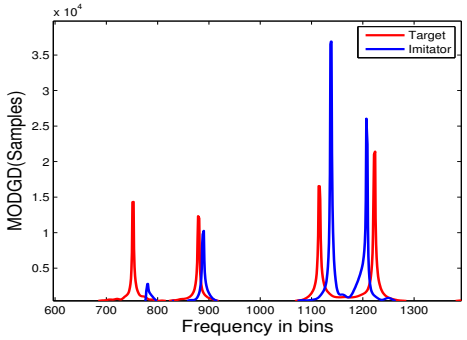

Figure 6: *Modified group delay functions computed for one frame in [p] for both target and imitator*

Table 1: *Accuracy in Top X-hit rate (%)*

	Top-X rate	
Method	**X=1**	**X=2**
i_{MFCC}	40.90	68.18
i_{MODGDF}	31.81	63.63
i_{MFCC} + i_{MODGDF}	45.45	72.72
i_{MFCC} • i_{MODGDF}	40.90	72.72
i_{MFCC} + i_{MODGDF} + P (Feat Fusion)	**50.00**	**81.81**

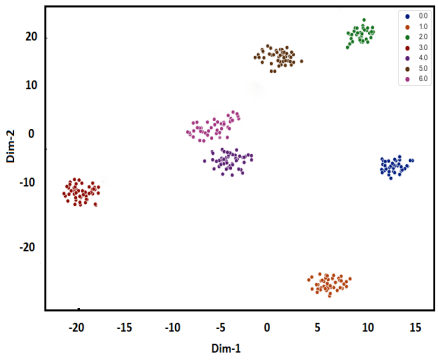

Figure 7: *t-SNE plots for the output vectors of the hidden layer for seven targets (different colours).*

5. Conclusion

The competency evaluation in voice mimicking is analyzed using i-vector framework and its fusion with prosodic features in the proposed work. i-vectors are computed from MFCC and MODGDF. Prosodic features are computed from the Legendre polynomial base. The experiment is carried out with various fusion scheme. DNN classifier is trained using a feature set computed from the target voices of 22 celebrities and tested with mimicry voices of professional mimicry artists. The system gives a top-2 hit rate of 81.81% for the fusion approach. Experimental results demonstrate that the i-vectors computed from Fourier transform magnitude and phase, with prosodic fusion have merit in evaluating the performance of voice imitators and related applications.

6. References

[1] L. Mary, A. Babu, A. Joseph, and G. M.George, "Evaluation of mimicked speech using prosodic features," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 7189–7193, May 2013.

[2] A. Eriksson and P. Wretling, "How flexible is the human voice?- a case study of mimicry," *Proc. of European Conference on Speech Technology*, pp. 1043–1046, 1997.

[3] E. Zetterholm, "Same speaker-different voices, a study of one impersonator and some of his different imitations," *Proc. of International Confernce on Speech Science and Technology*, pp. 70 –75, 2006.

[4] M. Farrus, M. Wagner, J. Anguita, and J. Hernando, "Robustness of prosodic features to voice imitation," *Proc. of Interspeech*, pp. 613 –616, September 2008.

[5] D.Gomati, S. A. Thati, K. V. Sridaran, and B. Yegnanarayana, "Analysis of mimicry speech," *Proc. of Interspeech*, pp. 813 –816, September 2008.

[6] E. Zetterholm, M. Blomberg, and D. Elenius, "The impact of semantic expectation on the acceptance of a voice imitation," *Proc. of the 9th Australian International Conference on speech Science and Technology,*, pp. 379–384, December 2002.

[7] T. Kitamura, "Acoustic analysis of imitated voice produced by a professional impersonator," *Proc. of Interspeech*, pp. 813 –816, September 2008.

[8] A. Diment, P. Rajan, T. Heittola, and T. Virtanen, "Modified group delay feature for musical instrument recognition," *Proc. of 10th International Symposium Computer Music Multidisciplinary Research, Marseille, France*, pp. 431–438, 2013.

[9] K. K. Paliwal and L. D. Alsteris, "On the usefulness of STFT phase spectrum in human listening tests," *Speech Communication*, vol. 45, pp. 153–170, 2005.

[10] G. Shi, M. M. Shanechi, and P. Aarabi, "On the importance of phase in human speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1867–1874, 2006.

[11] H. A. Murthy and B. Yegnanarayana, "Group delay functions and its application to speech processing," *Sadhana*, vol. 36, no. 5, pp. 745–782, November 2011.

[12] ——, "Formant extraction from minimum phase group delay function," *Speech Communication*, vol. 10, pp. 209–221, August 1991.

[13] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "i-vector based speaker recognition on short utterances," *Proc. of Interspeech*, pp. 2341–2344, 2011.

[14] J. Zhong, W. Hu, F. Soong, and H. Meng, "DNN i-vector speaker verification with short, text-constrained test utterances," *Proc. of Interspeech*, pp. 1507–1511, 08 2017.

[15] G. Ashour and I. Gath, "Characterization of speech during imitation," *Proc. of Eurospeech*, September 1999.

[16] R. Rajan and H. A. Murthy, "Group delay based melody monopitch extraction from music," *Proc. of the IEEE Int.Conf. on Audio, Speech and Signal Processing*, pp. 186–190, May 2013.

[17] ——, "Two-pitch tracking in co-channel speech using modified group delay functions," *Speech Communication*, vol. 89, pp. 37 – 46, 2017.

[18] A. V. Oppenheim and R. W. Schafer, *Discrete Time Signal Processing*. New Jersey: Prentice Hall, Inc, 1990.

[19] H. A. Murthy, "Algorithms for Processing Fourier Transform Phase of Signals," PhD Dissertation, Indian Institute of Technology, Department of Computer Science and Engg., Madras, India, December 1991.

[20] Hegde, Rajesh M., "Fourier transform based features for speech recognition," PhD Dissertation, Indian Institute of Technology Madras, Department of Computer Science and Engg., Madras, India, July 2005.

[21] N. Dehak, P. 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, pp. 788–798, 2011.

[22] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on Speech and Audio Processing*, vol. 13, pp. 345–354, 2005.

[23] C.-Y. Lin and H.-C. Wang, "Language identification using pitch contour information," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 601–604, 2005.

[24] E. Grabe, G. Kochanski, and J. Coleman, "Quantitative modelling of intonational variation," *Proc. of Speech Analysis and Recognition in Technology, Linguistics and Medicine*, 2003.

[25] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech,and Language Processing,*, vol. 15, no. 7, pp. 2095–2103, 2007.

[26] J. R. Orozco-Arroyave, J. C. Vásquez-Correa, and *et al.*, "Neurospeech: An open-source software for parkinson's speech analysis," *Digital Signal Processing*, 2017.

[27] J.-F. Bonastre, F. Wils, and S. Meignier, "Aliźe, a free toolkit for speaker recognition," *Proc. of Interspeech*, vol. 1, pp. 737–740, 2005.

[28] M. Ryynanen and A. Klapuri, "Query by humming of midi and audio using locality sensitive hashing," *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 2249–2252, 2008.

[29] T. Kinnunen and B. H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, pp. 12–40, 2010.

[30] M. Kotti, V. Moschou, and C. Kotropoulos, "Speaker segmentation and clustering," *Signal Processing*, vol. 88, no. 5, pp. 1091–1124, 2008.