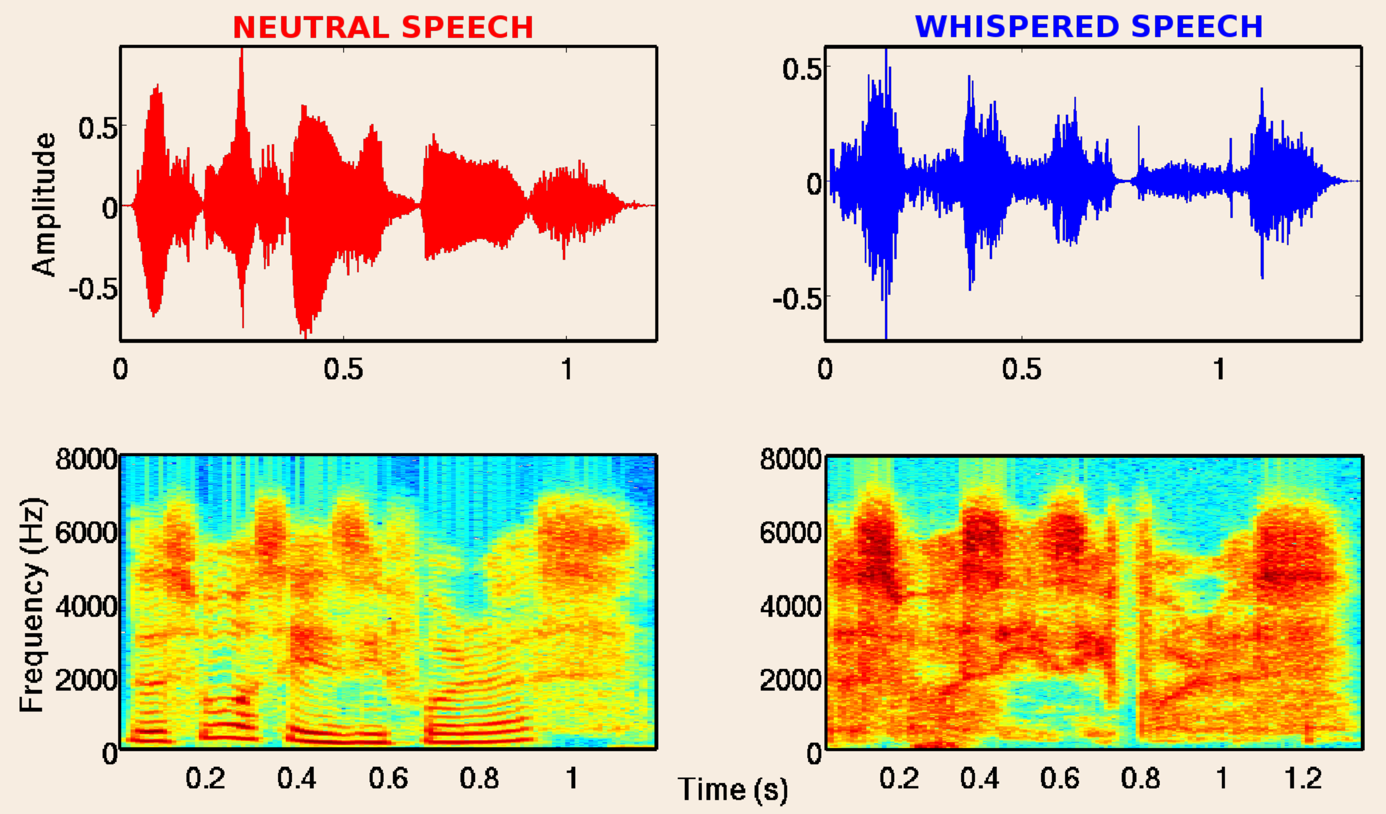
FORMANT-GAPS FEATURES FOR SPEAKER VERIFICATION USING WHISPERED SPEECH

Abinay Reddy Naini, Achuth Rao M V and Prasanta Kumar Ghosh

Electrical Engineering, Indian Institute of Science (IISc), Bangalore, India-560 012

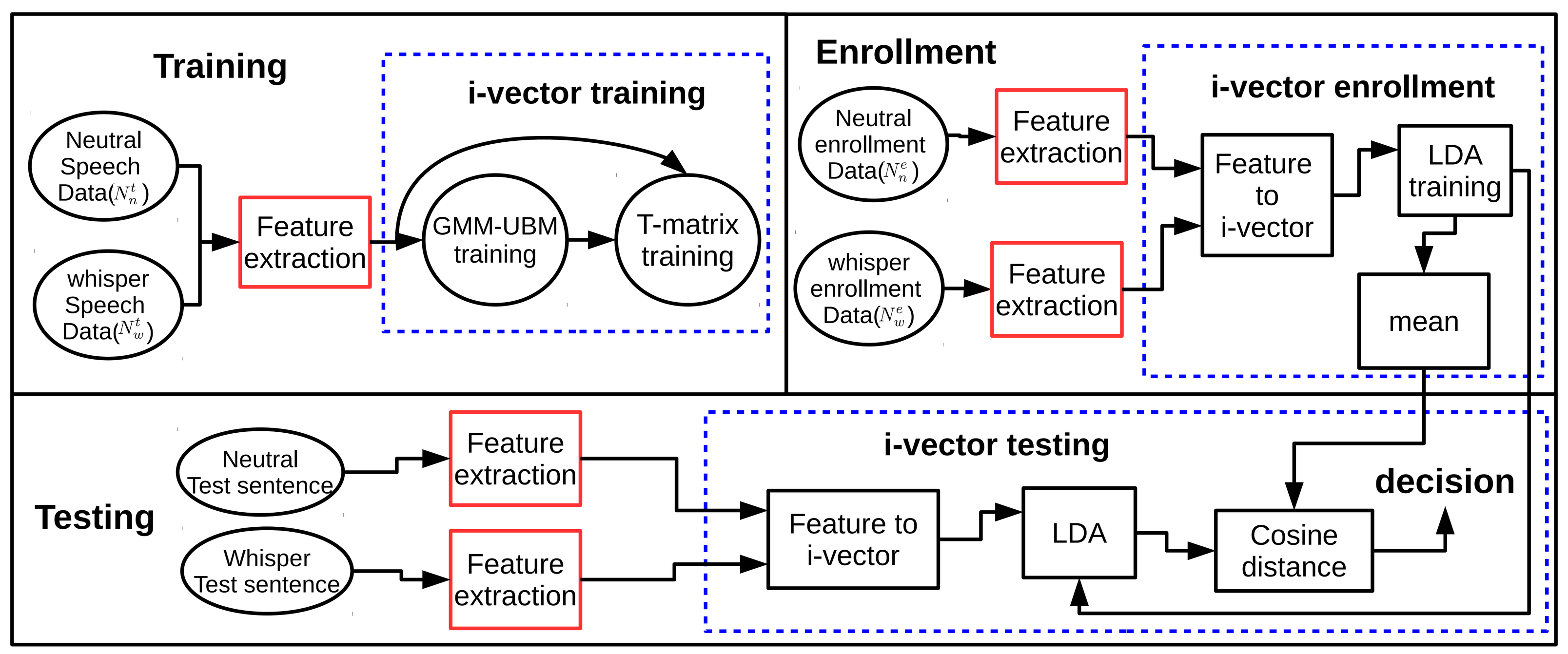
Introduction

- **Speaker verification(SV)**: To verify whether a given test speech recording is from an enrolled speaker or not.
- **Whisper speech**: Used in private conversations, pathological conditions.



- **Need for whisper SV**: Speakers often whisper the password in a biometric system, criminals might whisper in phone to avoid leaving the voice print[1].
- **Challenges**: Absence of pitch, Low-frequency formant shift, hyper-articulation

Whispered speaker verification system



- 3 major steps:[2]
  1) **Training**: GMM based background model and T-matrix training using available neutral and whisper training data.
  2) **Enrollment**: Involves extracting i-vectors using available neutral and whisper data of enrolled speakers.
  3) **Testing**: Taking decision using cosine distance between test speech i-vector and enrolled speaker i-vector.

Proposed Formant-Gaps features

- For each frame, we computed five formants using [3], 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. Let us consider first (f_i^1) and second order (f_i^2) formant gaps (FoGs) as

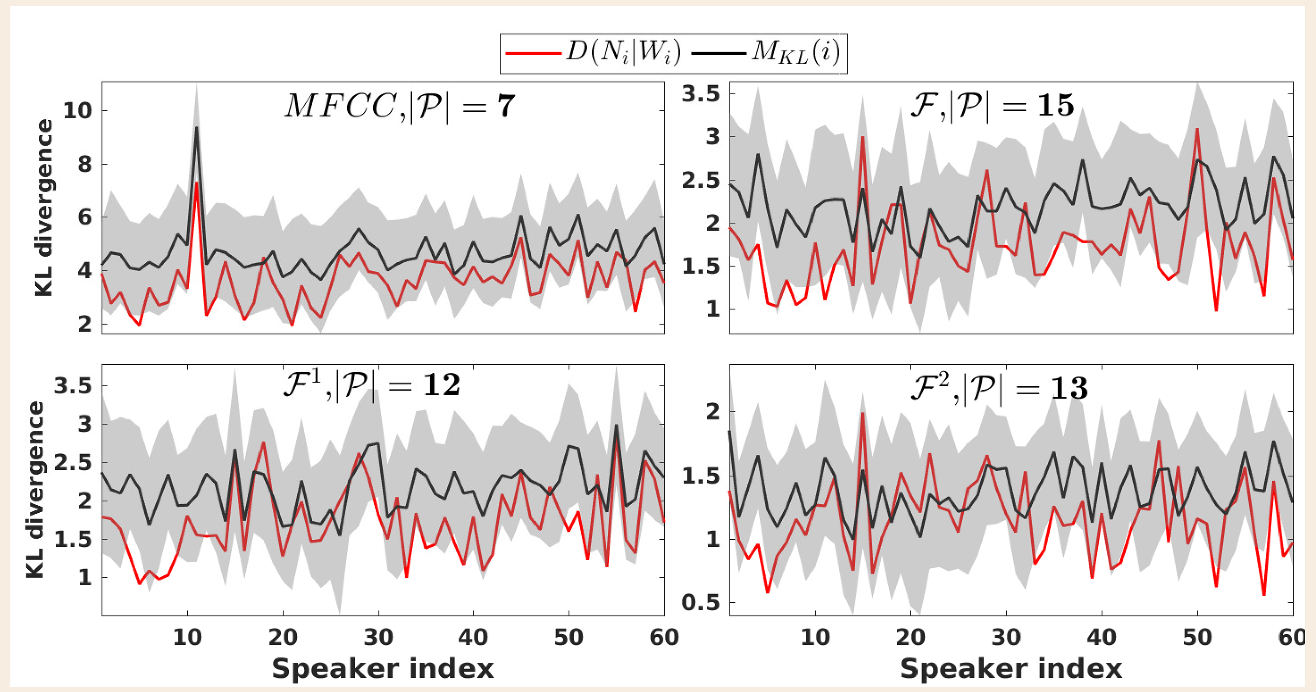
$$f_i^1 = f_{i+1} - f_i,\ 1 \le i \le 4, \qquad f_i^2 = f_{i+1}^1 - f_i^1,\ 1 \le i \le 3 \qquad (1)$$

Let $\mathcal{F}^1 = \{f_i^1; 1 \le i \le 4\}$, $\mathcal{F}^2 = \{f_i^2; 1 \le i \le 3\}$.

- We experimented two features using FoGs, namely,

$$FoG_1 = [\mathcal{F}, \mathcal{F}^1] \quad \text{and} \quad FoG_2 = [\mathcal{F}, \mathcal{F}^1, \mathcal{F}^2].$$

  - where the dimension of features FoG_1,FoG_2 are 9,12 respectively.
- **Illustrative experiment**:



In order to understand the distribution of the proposed features, we trained a speaker specific GMM for whispered and neutral speech features separately.

- $D(N_i|W_i)$: The KL divergence between i-th speaker's neutral GMM (N_i) and whispered GMM (W_i).
- $M_{KL}(i)$: The average of KL divergence between the N_i and $W_{j\neq i}$ speakers.

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

where $\mathcal{P} = \{i : D(N_i|W_i) < M_{KL}(i) - 1.5 \times \sigma_{KL}(i)\}$.

References

- [1] X Fan and J HL Hansen, Speaker identification within whispered speech audio streams,IEEE transactions on audio, speech, and language processing, vol.19, no.5, pp.14081421, 2011.
- [2] N Dehak, P J Kenny, R Dehak, P Dumouchel, and P Ouel-let, Front-end factor analysis for speaker verification, IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788798, 2011.
- [3] B Bozkurt, T Dutoit, B Doval, and C dAlessandro, Improved differential phase spectrum processing for formant tracking, in Eighth International Conference on Spoken Language Processing, 2004.
- [4] 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.437456, 2017.

Experiments & Results

- **Data set**: We considered data from 3 databases (CHAINS,wTIMIT,TIMIT) with 714 speakers comprising 29232 neutral and 22932 whispered recordings.
- **Baseline features**:
  - **MFCC:** 13-dimensional mel frequency cepstral coefficients along with veclocity and acceleration coefficients to make 39 dimensional features.
  - **AAMF:** Auditory-inspired amplitude modulation features (40-dimensional)[4].
  - **DNN:** Deep neural network(DNN) based feature mapping on both MFCC and AAMF features are considered.
- **Equal error rate(EER) for different test conditions**:

Table: EER using proposed and baseline features

		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: EER with varying number of whisper recordings (N_w^e) in enrollment

N_w^e	AAMF_{DNN}	FoG_1
0	17.01	**13.00**
2	14.14	**10.82**
4	**8.61**	9.68
6	**6.14**	8.88
8	**4.78**	8.46

- The combination of $\mathcal{F}$ and $\mathcal{F}_1$ features (FoG_1) performs the best, when only neutral data used in enrollment and tested using whispered speech.
- The feature mapping on the baseline feature ($MFCC_{DNN}$ and $AAMF_{DNN}$) performs better compared to (MFCC and AAMF), when when only neutral data used in enrollement and tested using whispered speech.
- The SV using baseline features requires at least four whisper recordings in the enrollment phase for it to perform better than the proposed features.

Conclusion

- We proposed formant-gaps based features for whispered speaker verification. The experiments revealed that the proposed features are robust to the modes (whisper and neutral) of speech for SV applications.
- **Future work** : Experimeting with different feature mapping methods for whispered speaker verification.

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

nainireddy@iisc.ac.in, achuthr@iisc.ac.in, prasantg@iisc.ac.in