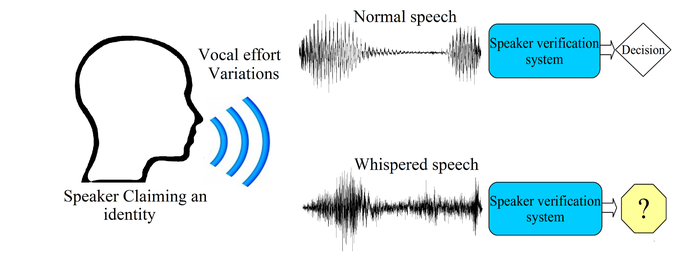
The effects of whispered speech on state-of-the-art voice based biometrics systems

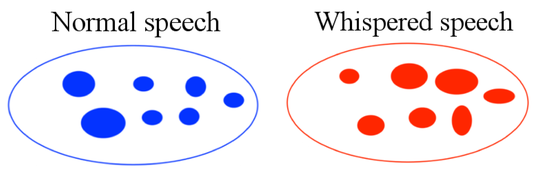
Milton Sarria-Paja and Tiago H. Falk

Institut National de la Recherche Scientifique (INRS-EMT), University of Quebec, Montréal, Quebec, Canada

INTRODUCTION

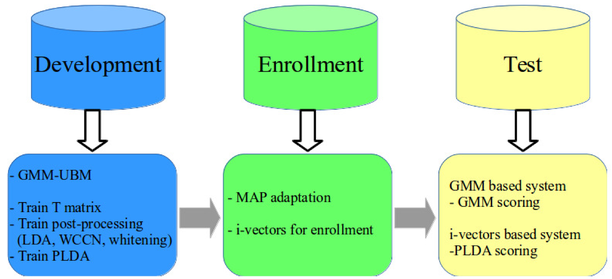


Mismatch between the training data and what the model encounters in real life.



SPEAKER VERIFICATION

Standard scheme



- Changes in the vocal effort have proven to affect significantly the performance of automatic speech processing systems.
- Whispered speech despite its reduced perceptibility, has been shown to convey relevant speaker identity and gender information.

Goal: Improve system performance for whispered speech without affecting performance for normal speech.

Quantify the effects of whispered speech on a standard SV system.

EXPERIMENTAL SETUP

Databases

Only normal speech

- **TIMIT database:** 630 speakers (438 Male, 192 Female).

Whispered and normal speech

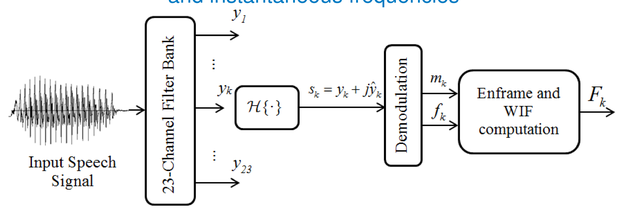
- **wTIMIT database:** 48 speakers (24 Male, 24 Female).
- **CHAINS Speech Corpus:** 36 speakers (20 Male, 16 Female).

Development: 476 speakers (462 from TIMIT, 14 from wTIMIT), Clients: 160 speakers (100 from TIMIT, 36 from CHAINS, 24 from wTIMIT).

Features

- Mel-frequency cepstral coefficients (MFCC)
- Weighted Instantaneous Frequencies (WIFs)

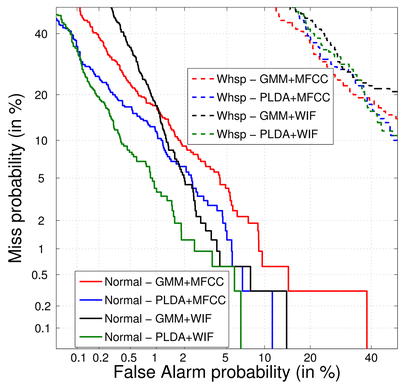
Speech signal decomposition in bandpass channels for estimation of envelope and instantaneous frequencies



BASELINE RESULTS

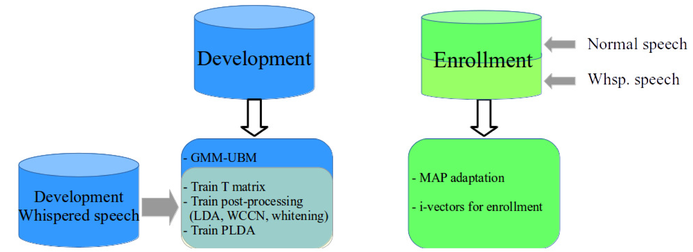
Train/Test mismatch effects

SV System	MFCC Norm	MFCC Whsp	WIF Norm	WIF Whsp
GMM (Map adaptation)	4.38	25.83	2.50	29.17
PLDA	3.13	27.50	1.60	26.35



From the DET curves and tabulated EER values, it can be observed that significant performance degradation occurs in mismatch conditions. There is a gap in performance between normal and whispered speech higher than 20% for all cases.

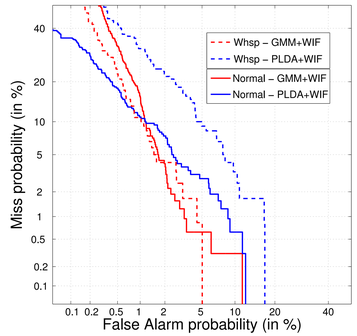
HOW TO ADDRESS THIS PROBLEM?



- Case 1: Include whispered speech during *T* matrix estimation.
- Case 2: Include whispered speech during enrollment.
- Case 3: Combination of the two cases above.

RESULTS

	MFCC Norm	MFCC Whsp	WIF Norm	WIF Whsp
GMM + MAP adaptation				
Case 2	6.01	10.00	2.19	2.70
PLDA based system				
Case 1	4.06	19.15	1.86	17.68
Case 2	5.63	10.35	4.00	8.40
Case 3	6.56	7.77	4.09	5.85



- GMM based system: Combined with AM-FM based features and in presence of whispered speech it seems to be more robust and is able to maintain the error rate for the two speaking styles below 3%.
- PLDA + i-vector based system: Highly sensitive to the addition of new data even if whispered speech features were included during total variability matrix estimation.

CONCLUSION

Addition of whispered speech during training and enrollment seems to highly affect current state-of-the-art SV systems. For the classical GMM based system that was not the case when combined with the use of AM-FM based features. Such finding suggests that the phase and envelope of bandpass signals can contain highly discriminative speaker specific information during normal and whispered speech. .

Institut National de la Recherche Scientifique (INRS-EMT), University of Quebec , Montréal, Quebec, Canada.

e-mail: sarria@emt.inrs.ca