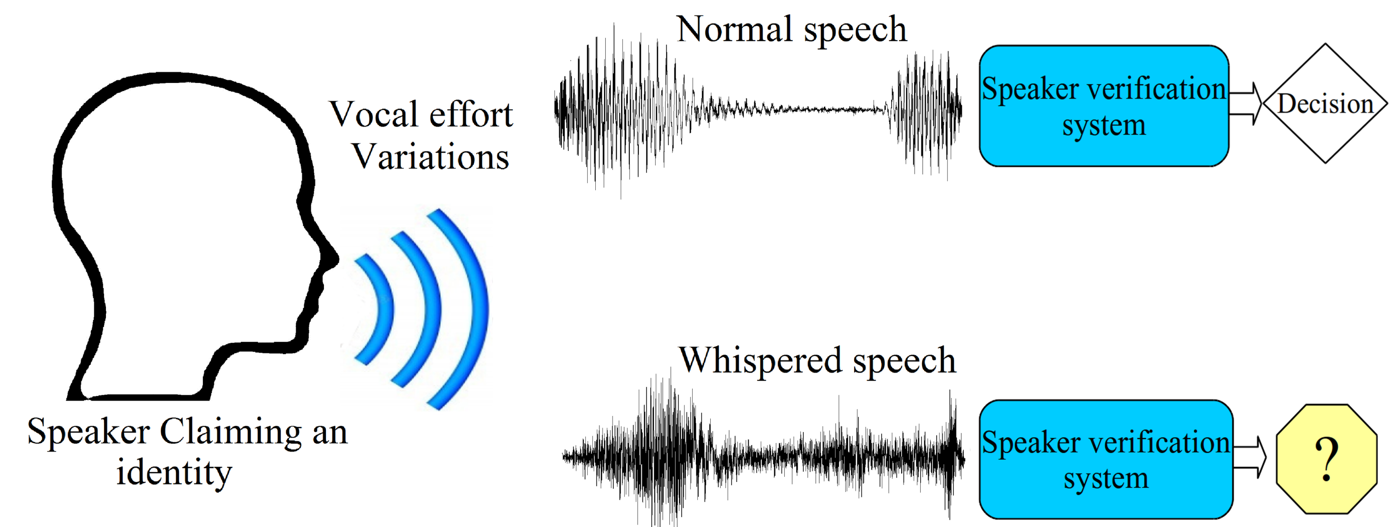
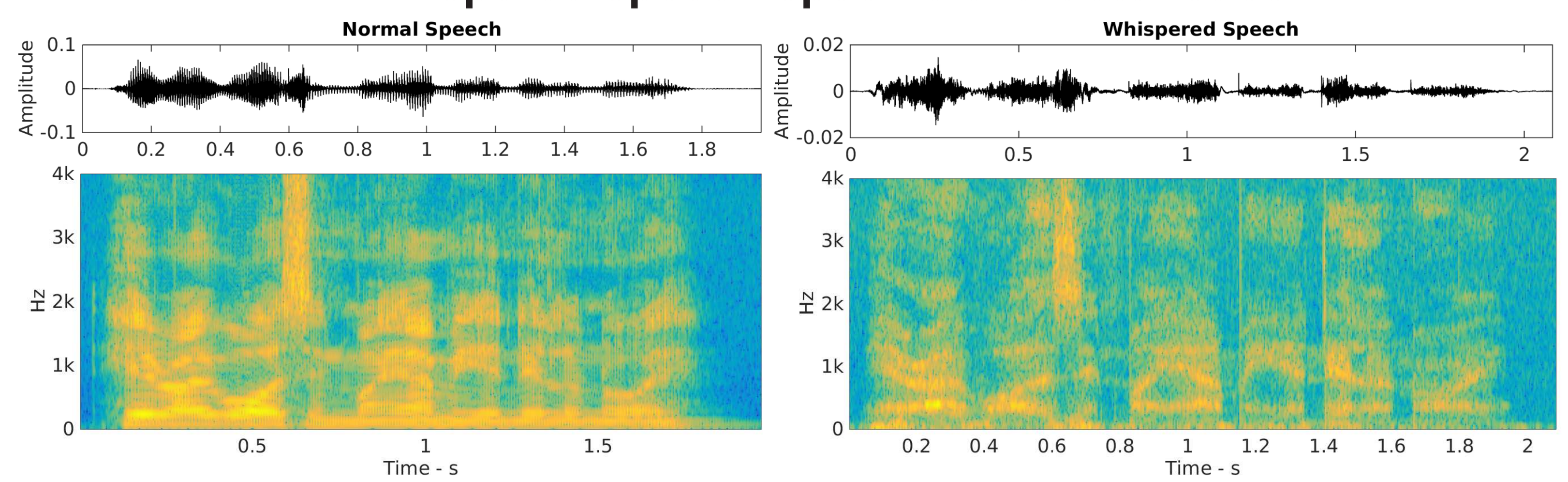
Feature Mapping, Score-, and Feature-Level Fusion for Improved Normal and Whispered Speech Speaker Verification

Milton Sarria-Paja, Mohammed Senoussaoui, Douglas O'Shaughnessy and Tiago H. Falk
Institut National de la Recherche Scientifique (INRS-EMT), University of Quebec, Montréal (QC) - Canada

INTRODUCTION



Whispered speech specific characteristics



- Lack of fundamental frequency.
- Formant shifts towards higher frequencies.
- Lower and flatter power spectral density.
- 64 *low level descriptors* (LLDs): spectral, prosody and voice quality were compared and 56 showed to be statistically different.

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

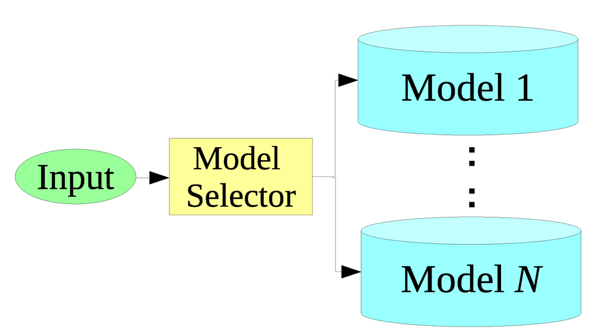
HOW TO ADDRESS THIS PROBLEM?

Three approaches have shown to be useful in related areas:

1) **Feature mapping:**
   - Compensate for the lack of data during enrollment
   - Compensate for the differences during testing

2) **Multiple model recognizer:**

Requires significant amounts of data to train the models.



3) **Multi-style models:**
During parameter estimation and enrollment combination of normal speech and small amounts of speech of varying vocal efforts is used.

EXPERIMENTAL SETUP

Databases

Database	Num. of speakers		recordings/speaker	
	Female	Male	Norm.	Whsp.
TIMIT	192	438	10	–
wTIMIT	24	24	450	450
CHAINS	16	20	37	37

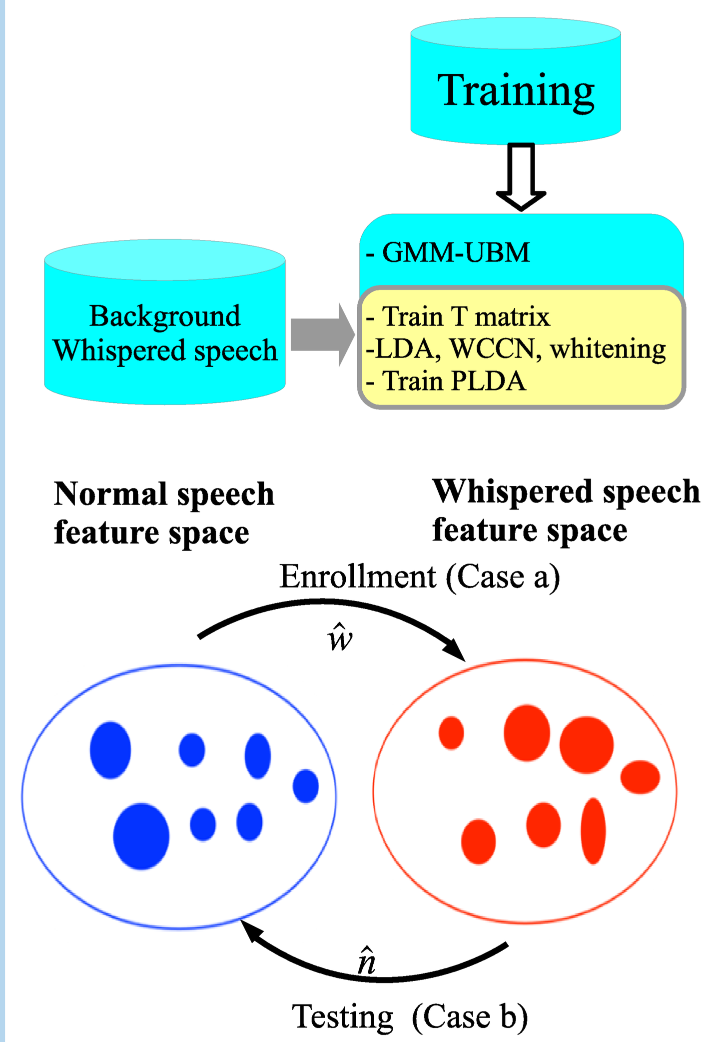
Task design

	Num. of speakers/Database			Total record.	
	TIMIT	wTIMIT	CHAINS	norm	whsp
UBM estimation	462	0	0	3696	0
T matrix estimation	462	14	0	9996	6300
Enrollment	100	24	36	1280	480
Testing	100	24	36	320	120

SV system parameters:
PLDA/i-vectors based system with:
UBM $C = \{64, 128, 256, 512\}$. T matrix $D = \{200, 300, 400\}$.

FEATURE MAPPING

Two approaches are evaluated: DNN and GMM based mappings



Mean Cepstral Distance and Root Mean Square Error

Evaluation Measures	Norm to Whsp		Whsp to Norm	
	GMM	DNN	GMM	DNN
MCD	13.84	12.78	13.96	12.75
ε_{rms}	0.644	0.596	0.649	0.595

EER comparison with the baseline system

Scenario	Norm	Whsp
Baseline	3.13	27.35
Whsp in dev. set	**4.06**	**19.15**

	Feature Mapping			
	GMM	DNN	GMM	DNN
Case a	8.75	6.25	24.17	20.00
Case b	**4.06**	4.06	**17.50**	21.07

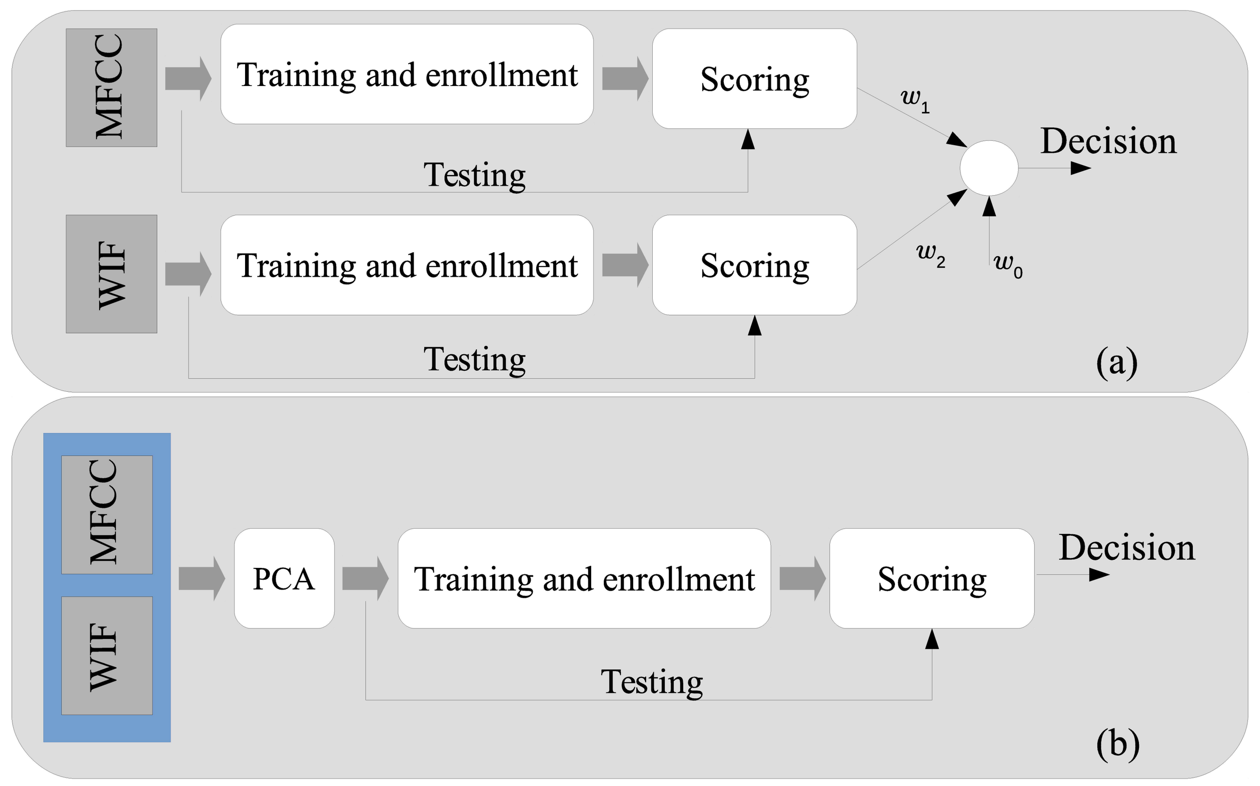
A direct mapping between whispered and normal speech features does not seem to help reducing error rates when testing with whispered speech (Except **Case b** using GMM). The mappings cannot transform effectively speaker specific characteristics associated with identity affected while whispering.

MULTI-STYLE MODELS

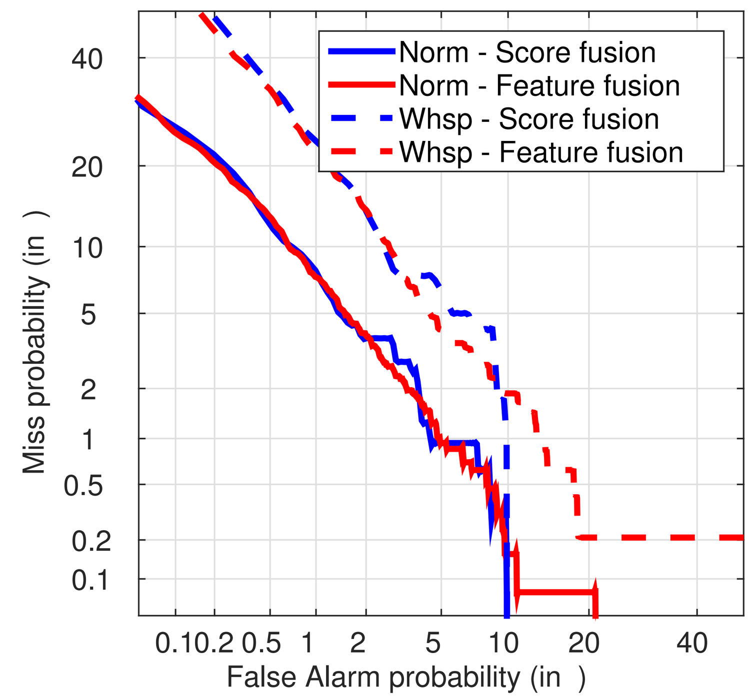
Addition of whispered speech during training and enrollment **comes with a cost**: ⇒

SV system	Norm	Whsp
Baseline/PLDA	2.93	28.00
Multy-Style/PLDA	5.56	8.90

To compensate for the losses:
Include complementary features. Explore fusion schemes: Fusion at the scoring level - SCF (a) and Fusion at the frame level - FF (b).



Scenario	Normal		Whisper	
	Fusion level			
	SCF	FF	SCF	FF
Case 1	1.56	0.74	15.49	17.69
Case 2	2.57	2.03	5.45	4.35



- **Case 1)** Whispered speech data only in training set.
- **Case 2)** Whispered speech data in training and enrollment sets.

CONCLUSION

- Two different approaches were compared in order to reduce error rates for SV with whispered speech while maintaining performance with normal speech.
- Multi-style models are the least computationally expensive and most effective way to achieve significant error rate reductions.
- Our approach to compensate multi-style models is to include AM-FM based features and use fusion schemes at the frame level and at the scoring level.
- Finally, it is observed that features that rely on instantaneous phase information add complementary speaker identity information.

The authors acknowledge funding from: NSERC - Canada, COLCIENCIAS - Colombia and SYTACom - Quebec, Canada.

e-mail: sarria@emt.inrs.ca