A PARAMETRIC APPROACH FOR CLASSIFICATION OF DISTORTIONS IN PATHOLOGICAL VOICES

AUDIO ANALYSIS LAB

Amir Hossein Poorjam [1], Max A. Little [2,3], Jesper Rindom Jensen [1], Mads Græsbøll Christensen [1]

[1] Audio Analysis Lab, CREATE, Aalborg University, DK [2] Engineering and Applied Science, Aston University, Birmingham, UK [3] Media Lab, MIT, Cambridge, Massachusetts, USA

[1] {ahp, jrj, mgc}@create.aau.dk, [2] max.little@aston.ac.uk

AALBORG UNIVERSITY DENMARK

Introduction

- Information on the type of distortion corrupting a signal can be used to inform the choice of appropriate enhancement algorithms.
- Most existing methods focused on detecting a single and specific type of distortion in a signal.
- In [1], we proposed a method to classify four major types of distortion in vowels directly from MFCCs extracted from speech signals.
- Limitations of [1]:
  - MFCCs encode not only distortion in signals, but also other variability (speaker, articulation and disorder).
  - Distortion classification decision is made by majority vote over all frames, and the computation time increases with increasing signal length.
- In this paper, distortion in variable duration recordings is modeled with a fixed-length, low-dimensional vector.

Distortion Modeling

- **Channel variability** can be produced artificially by corrupting the clean recording by different types and levels of distortion.
- **Method:**
  - Fitting a Gaussian mixture model (GMM) to the features of a recording.
  - Assuming that the GMM mean supervector of the r^{th} recording from the s^{th} speaker can be decomposed as:

$$\boldsymbol{M}_{s,r} = \boldsymbol{m} + \boldsymbol{V}\boldsymbol{y}_s + \boldsymbol{U}\boldsymbol{x}_{s,r} + \boldsymbol{D}\boldsymbol{z}_s \,. \qquad (1)$$

- **Definitions:**
  - $\boldsymbol{m}$ is speaker- and channel-independent supervector,
  - $\boldsymbol{V}$ is a rectangular matrix of low rank with high speaker variability
  - $\boldsymbol{y}_s$ is the speaker factor
  - $\boldsymbol{U}$ is a rectangular matrix of low rank with high channel variability
  - $\boldsymbol{x}_{s,r}$ is the channel factor containing channel related information
  - $\boldsymbol{D}$ is a diagonal matrix describing any remaining speaker variability
  - $\boldsymbol{z}_s$ is the speaker-specific residual factor
  - The factors $\boldsymbol{x}_{s,r}$, $\boldsymbol{y}_s$ and $\boldsymbol{z}_s$ are assumed to be independent of each other and have a standard normal prior distribution.
- **Estimating the matrices $\boldsymbol{V}$, $\boldsymbol{U}$, $\boldsymbol{D}$, and the vectors $\boldsymbol{x}_{s,r}$, $\boldsymbol{y}_s$ and $\boldsymbol{z}_s$ [2]:**
  1) Train $\boldsymbol{V}$, assuming that $\boldsymbol{U}$ and $\boldsymbol{D}$ are zero.
  2) Estimate $\boldsymbol{U}$ given the estimate of $\boldsymbol{V}$ and assuming that $\boldsymbol{D}$ is zero.
  3) Estimate the residual matrix $\boldsymbol{D}$ given the estimates of $\boldsymbol{V}$ and $\boldsymbol{U}$.
  4) $\boldsymbol{x}_{s,r}$, $\boldsymbol{y}_s$ and $\boldsymbol{z}_s$ are then calculated given the estimates of $\boldsymbol{V}$, $\boldsymbol{U}$ and $\boldsymbol{D}$.

Channel Factor and Subspace Estimation

- The channel factor $\boldsymbol{x}_{s,r} \sim N(\boldsymbol{\mu}_{s,r}, \boldsymbol{\Lambda}_{s,r})$ and the channel subspace $\boldsymbol{U}$ are estimated by applying an EM algorithm [2].
- **In the E-step**, using a random initialization of $\boldsymbol{U}$, the posterior distribution of the channel factor is calculated as:

$$\boldsymbol{\mu}_{s,r} = E[\boldsymbol{x}_{s,r}] = (\boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_s\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\mathbf{f}_{s,r} \qquad (2)$$

$$\boldsymbol{\Lambda}_{s,r} = E[\boldsymbol{x}_{s,r}\boldsymbol{x}_{s,r}^T] = \boldsymbol{\mu}_{s,r}\boldsymbol{\mu}_{s,r}^T + (\boldsymbol{I} + \boldsymbol{U}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{N}_s\boldsymbol{U})^{-1}. \qquad (3)$$

- **In the M-step**, the channel subspace is updated by solving the equations:

$$\boldsymbol{U}_i\boldsymbol{\Theta}_c = \boldsymbol{\Psi}_i \,. \qquad (4)$$

- **Definitions:**
  - $\boldsymbol{\Sigma}$ is a block-diagonal matrix entries form the covariance matrix of the c^{th} mixture of the UBM,
  - $N_{s,r,c} = \sum_{l=1}^{L}\gamma_{c,l}$ and $\mathbf{f}_{s,r,c} = \sum_{l=1}^{L}\gamma_{c,l}[\boldsymbol{\rho}_l - (\boldsymbol{m}_c + \boldsymbol{V}_c\boldsymbol{y}_s)]$ are the zero- and first order statistics for each speaker s, recording r and mixture component c.
  - $\boldsymbol{\rho}_l$ is the acoustic features of the l^{th} frame
  - $\boldsymbol{I}$ is an identity matrix,
  - $\boldsymbol{N}_s$ is a block-diagonal matrix which its entries are $(\sum_r N_{s,r,c})\boldsymbol{I}$
  - $\mathbf{f}_{s,r}$ is a vector constructed by concatenation of $\mathbf{f}_{s,r,c}$
  - $\gamma_{c,l}$ is the posterior probability of the c^{th} mixture generating $\boldsymbol{\rho}_l$,
  - $\boldsymbol{m}_c$ and $\boldsymbol{V}_c$ are, respectively, the subvector of $\boldsymbol{m}$ and the submatrix of $\boldsymbol{V}$ of mixture component c.
  - $\boldsymbol{\Theta}_c = \sum_s\sum_r N_{s,r,c}\boldsymbol{\Lambda}_{s,r} \quad c = 1, \ldots, C$
  - $\boldsymbol{\Psi}_i$ is the i^{th} row of $\boldsymbol{\Psi} = \sum_s\sum_r \mathbf{f}_{s,r,c}\boldsymbol{\mu}_{s,r}^T$

The Proposed Method

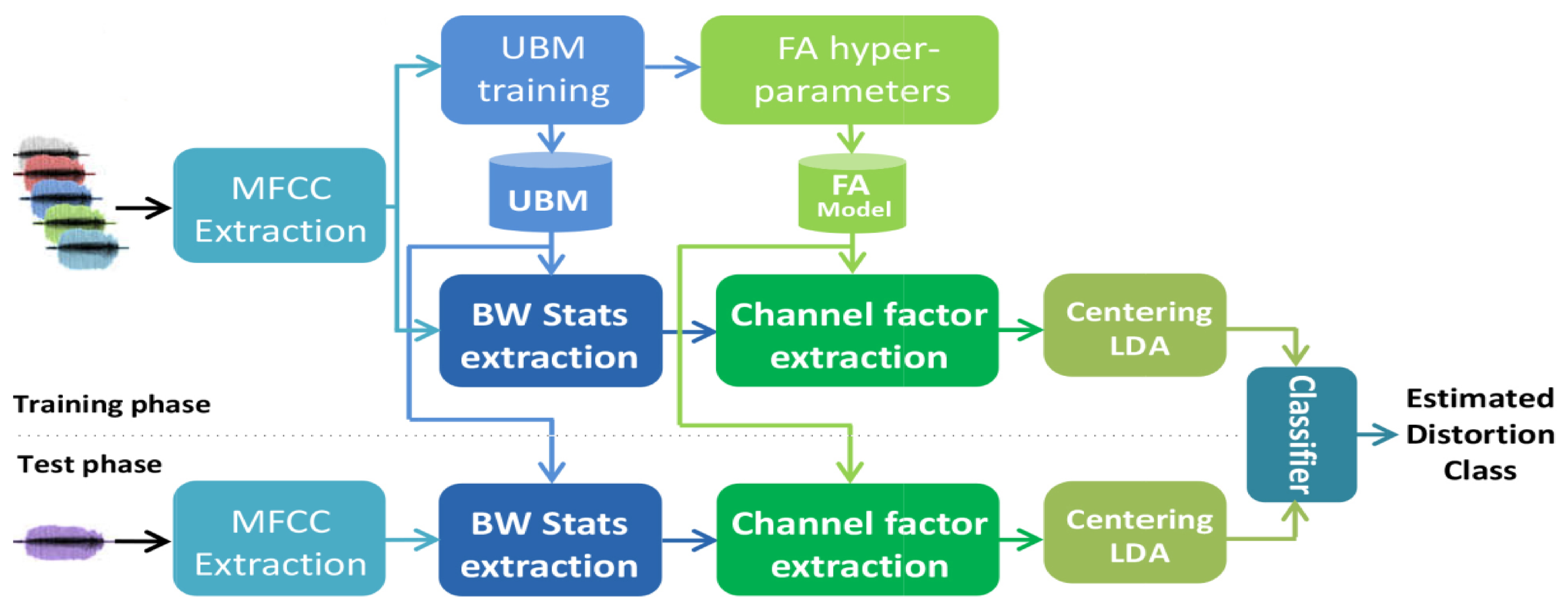

Fig. 1: The block diagram of the proposed distortion classification system.

Experimental Setup

- **Database:**
  - Parkinson's voice database (sustained vowels, 750 telephone recordings).
- **Distortion Classes:**
  - Additive noise (white Gaussian, babble, office ambiance noises)
  - Reverberation (8 different real room impulse responses)
  - Peak clipping (clipping level: 0.3, 0.4, 0.5, 0.6)
  - Coding (6.3 kbps, 9.6 kbps and 16 kbps CELP codecs)
- **Acoustic features:**
  - 39 dimensional vector (12 MFCCs + frame energy + Δ + ΔΔ)
- **Distortion Modeling:**
  - GMM with 256 mixtures
  - Speaker factor dim.: 0
  - Channel factor dim.: 210
- **Classifiers:**
  - SVM with RBF kernel
  - PLDA

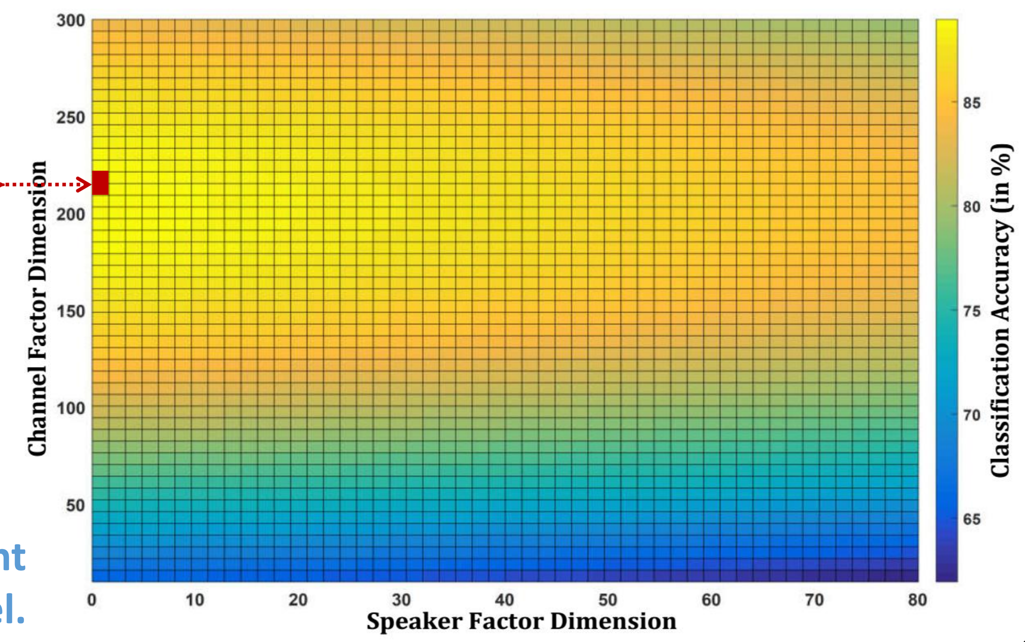

Fig. 2: Performance of different configuration of the FA model.

Results

Table 1: Comparison of [1] and the proposed method before and after pre-processing channel vectors using LDA. Results are in the form of mean ± STD.

System	Clean	Noisy	Reverb.	Clipped	Coded	Overall
Baseline	55 ± 11	97 ± 4	77 ± 4	82 ± 7	85 ± 9	79 ± 3
PLDA	100 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	20 ± 0
PLDA + LDA	77 ± 4	98 ± 2	86 ± 4	82 ± 2	93 ± 3	87 ± 1
SVM	28 ± 18	33 ± 5	31 ± 16	35 ± 14	68 ± 12	39 ± 4
SVM + LDA	78 ± 3	97 ± 2	87 ± 4	85 ± 2	93 ± 3	88 ± 1

Conclusions

- Distortion in variable duration signals is modeled by a fixed-length, low-dimensional vector which is more suitable for classification algorithms.
- Channel vectors are more robust to small changes in signal characteristics than MFCCs, they are more suitable for distortion classification in pathological voices.

References

[1] A. H. Poorjam, J. R. Jensen, M. A. Little, and M. G. Christensen, "Dominant distortion classification for pre-processing of vowels in remote biomedical voice analysis," in INTERSPEECH, 2017, pp. 289–293.

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A Study of inter-speaker variability in speaker verification," IEEE Trans. Audio. Speech. Lang. Processing, vol. 16, no. 5, pp. 980–988, 2008.