Feature Compensation Techniques for ASR on Band-Limited Speech

Author : Nicolás Morales, Doroteo Torre Toledano,
John H. L. Hansen, Javier Garrido

Professor: 陳嘉平

Reporter: 吳國豪

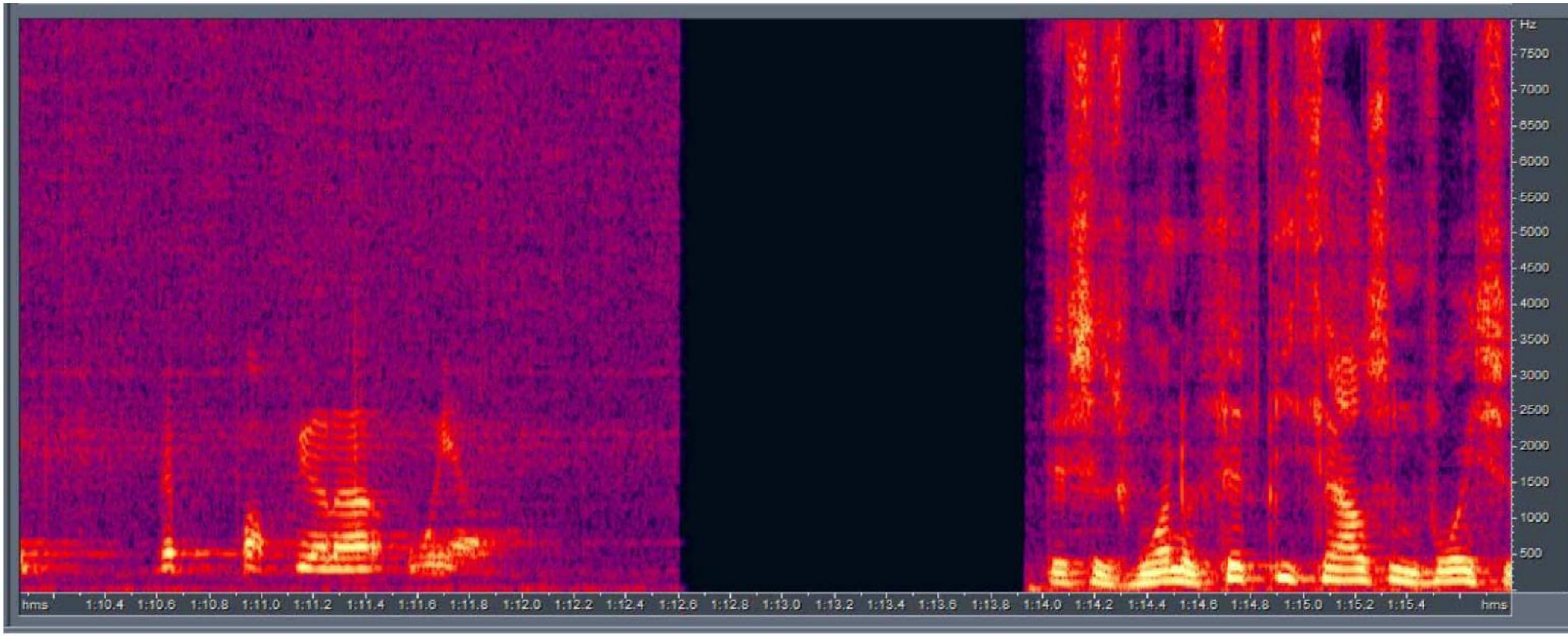
Outline

- Introduction
- Mathematical model of the effect of band-limiting distortions on MFCC features
- Feature Compensation
- Experiments

Introduction

- Band-limited speech: speech for which parts of the spectrum are completely lost

- **Band-limited speech** is a major cause for accuracy degradation of automatic speech recognition (ASR) systems particularly when acoustic models have been trained with data with a different spectral range.


Hz
7500
7000
6500
6000
5500
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
hms
1:10.4
1:10.6
1:10.8
1:11.0
1:11.2
1:11.4
1:11.6
1:11.8
1:12.0
1:12.2
1:12.4
1:12.6
1:12.8
1:13.0
1:13.2
1:13.4
1:13.6
1:13.8
1:14.0
1:14.2
1:14.4
1:14.6
1:14.8
1:15.0
1:15.2
1:15.4
hms

Introduction

- In this paper, we present an extensive study of the problem of ASR of band-limited speech with full-bandwidth acoustic models.

- Our focus is mainly on **band-limited feature compensation.**

Mathematical model

- The effect of a convolutional distortion may be expressed for the power spectrum as

$$\left|Y_t(f)\right|^2 = \left|H_t(f)\right|^2 \cdot \left|X_t(f)\right|^2$$

where $Y_t(f)$ and $X_t(f)$ represent the spectra for the distorted and original signals, respectively, for a time frame , and $H_t(f)$ is the frequency response of the distortion.

Mathematical model

- When the front-end employed is derived from a bank of filters, the following approximation is typically assumed for each filter. (j is the order of the filter, and f_j the center frequency of the filter)

$$\left|Y_t(f_j)\right|^2 \approx \left|H_t(f_j)\right|^2 \cdot \left|X_t(f_j)\right|^2 => \left|Y_t(f_j)\right|^2 \approx h_{j,t} \cdot \left|X_t(f_j)\right|^2$$

$$\left|Y_t(f_j)\right|^2 = h_{j,t} \cdot \left|X_t(f_j)\right|^2 + e_{j,t}, \text{ where} \begin{cases} h_{j,t} = 0, if \;\; j \in F \\ h_{j,t} = 1, if \;\; j \notin F \end{cases}$$

F represents the channels in the filterbank removed by the bandwidth-limitation

Mathematical model

- The general definition of MFCCs is

$$x_{i,t} = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} \log(\left| X_{j,t} \right|^2) \cos\left(\frac{\pi i}{N} (j - 0.5) \right)$$

where subindex *i* is the order of the MFCC coefficient, *t* represents a time frame, and *N* is the number of channels in the filterbank.

$$x_i = \sum_{j=1}^{N} C_{ij} \cdot \log(\left| X_j \right|^2)$$

$$C_{ij} = \sqrt{\frac{2}{N}} \cdot \cos\left(\frac{\pi i}{N} (j - 0.5) \right)$$

Mathematical model

- MFCC features of band-limited speech are

$$y_i = \sum_{j=1}^{N} C_{ij} \cdot \left(\log(h_j \cdot \left| X_j \right|^2 + e_j) \right)$$

- The difference between full-bandwidth and band-limited MFCC vectors for a particular frame

$$x_i - y_i = \sum_{j=1}^{N} C_{ij} \cdot \left[\log(\left| X_j \right|^2) - \log(h_j \cdot \left| X_j \right|^2 + e_j \right]$$

Mathematical model

- Now we decompose the sum over all filters in the filterbank into 2 terms corresponding to channels affected by the bandwidth restriction and intact channels, respectively

$$x_i - y_i = \sum_{j=1,\, j\notin F}^{N} C_{ij} \cdot \left[\log(\left|X_j\right|^2) - \log(h_j \cdot \left|X_j\right|^2 + e_j \right]$$
$$+ \sum_{j=1,\, j\in F}^{N} C_{ij} \cdot \left[\log(\left|X_j\right|^2) - \log(h_j \cdot \left|X_j\right|^2 + e_j \right]$$

Mathematical model

- We may then approximate full-bandwidth MFCCs as

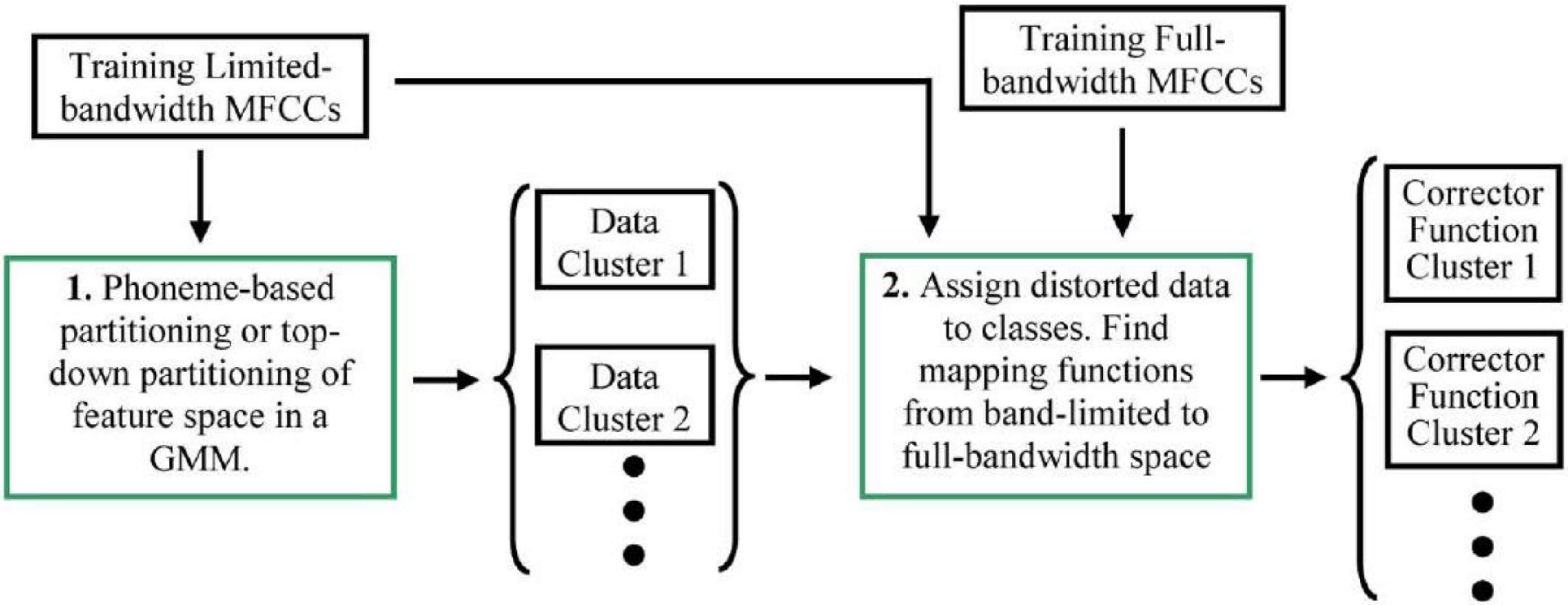
$$x_i \approx y_i + \sum_{j=1,\, j\in F}^{N} C_{ij} \cdot \left[\log(\left|X_j\right|^2) - \log(e_j) \right]$$

- In practice, values of e_j are random and significantly smaller than the values of the original signal

$$x_i \approx y_i + \sum_{j=1,\, j\in F}^{N} C_{ij} \cdot \left[\log(\left|X_j\right|^2) \right]$$

Full-Bandwidth Estimation

- **First**, training data is divided into clusters.
- **Second**, for each cluster a set of corrector functions is trained.
- **Third**, full-bandwidth features are estimated from band-limited data.


Training Limited-bandwidth MFCCs
Training Full-bandwidth MFCCs
1. Phoneme-based partitioning or top-down partitioning of feature space in a GMM.
Data Cluster 1
Data Cluster 2
2. Assign distorted data to classes. Find mapping functions from band-limited to full-bandwidth space
Corrector Function Cluster 1
Corrector Function Cluster 2

Full-Bandwidth Estimation

- **Partitioning** is initialized with a single cluster defined as a Gaussian distribution with mean $\mu_{0,0}$ and covariance $\Sigma_{0,0}$, computed for all training data.

- This initial cluster is divided into two by perturbing the mean vector by $\pm\eta$ times the vector of standard deviations, where η is a perturbation factor.

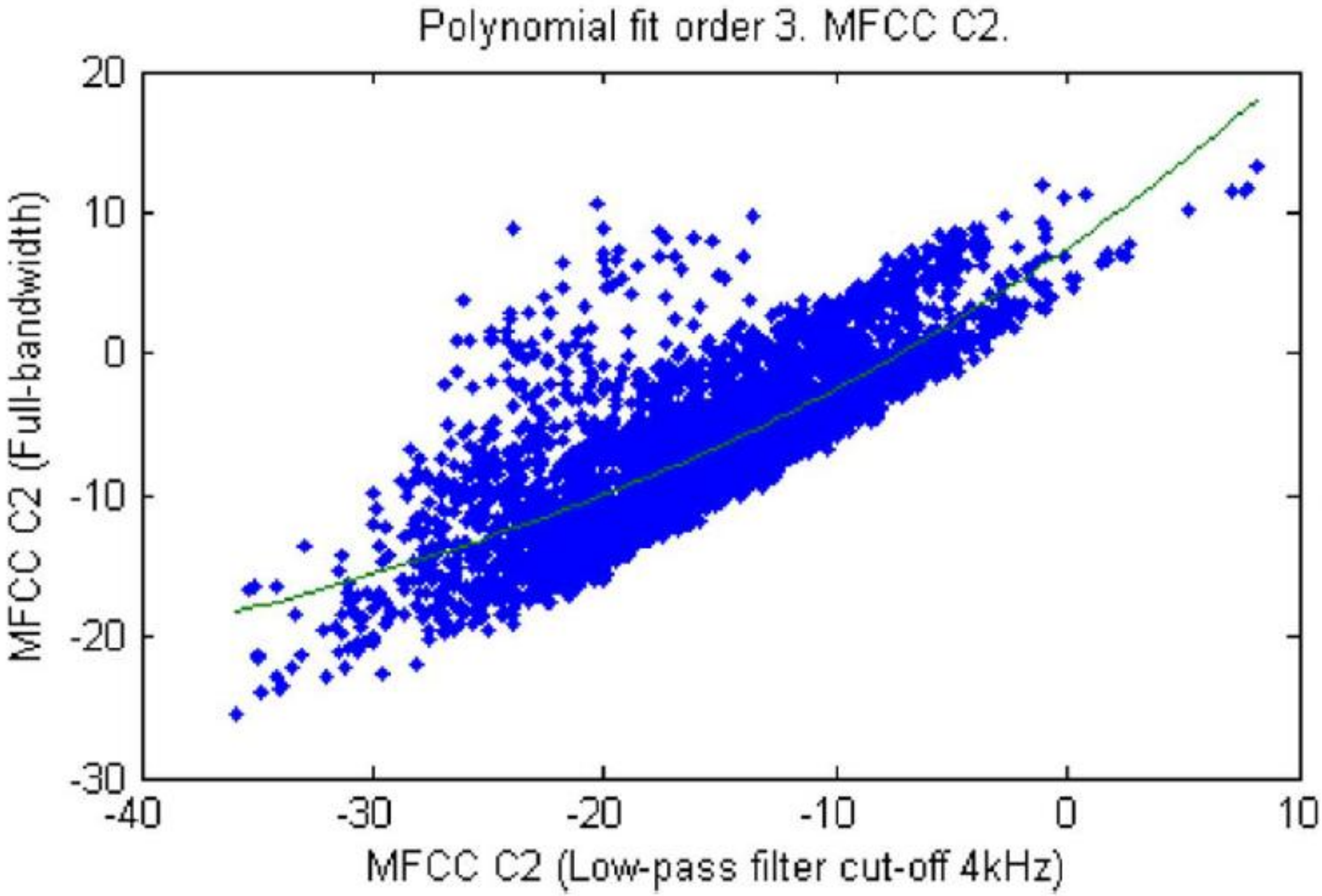
- Training data are reassigned to either cluster, and means and covariances are recalculated.

Full-Bandwidth Estimation

- Two methodologies are proposed depending on whether **stereo data** are available for training, or not.

- Training of Corrector Functions With **Stereo Data**:

  When stereo data are available, it is possible to learn a mapping from limited-bandwidth to full-bandwidth data using **linear least squares curve fitting techniques**.


Polynomial fit order 3. MFCC C2.
MFCC C2 (Full-bandwidth)
MFCC C2 (Low-pass filter cut-off 4kHz)
20
10
0
-10
-20
-30
-40
-30
-20
-10
0
10

Full-Bandwidth Estimation

- Training of Corrector Functions With **Nonstereo Data**:

- In the most general case, the vector of means and matrix of covariance of a cluster in the full-bandwidth feature space are related to their limited-bandwidth counterparts as

$$\mu_{x,k} = \mu_{y,k} + r_k$$

$$\Sigma_{x,k} = \Sigma_{y,k} + R_k$$

Full-Bandwidth Estimation

- The probability of observation of a feature vector in the full-bandwidth space is

$$p(x) = \sum_{k=1}^{K} N(x; \mu_{x,k}, \Sigma_{x,k}) \cdot p(k)$$

- Using an expectation-maximization (EM) strategy:

$$r_k = \frac{\sum_{t=1}^{T} p(k \mid x_{t,} \phi) \cdot x_t}{\sum_{t=1}^{T} p(k \mid x_{t,} \phi)} - \mu_{y,k} \qquad R_k = 0$$

Feature Compensation

- Thus, for the limited-bandwidth space the probability of observing a feature vector is

$$p(y) = \sum_{k=1}^{K} p(y \mid k) \cdot p(k) = \sum_{k=1}^{K} N(y; \mu_k, \Sigma_k) \cdot p(k)$$

- Assuming *x* and *y* jointly Gaussian within a cluster of data pairs *k* the conditional expectation of clean feature vectors given the distorted vectors and the cluster is

$$\begin{aligned} E\{x \mid y, k\} &= \mu_{x,k} + \Sigma_{xy,k} (\Sigma_{y,k})^{-1} (y - \mu_{y,k}) \\ &= B_k y + b_k \end{aligned}$$

We call B_k and b_k the compensation matrix and offset vector, respectively.

Feature Compensation

- **Estimation of undistorted features**:

  Using for example the minimum mean squared error (MMSE) criterion:

$$
\begin{aligned}
x^{MMSE} &= E\{x \mid y\} = \sum_{k=1}^{K} p(k \mid y) \cdot E\{x \mid y, k\} \\
&= \sum_{k=1}^{K} p(k \mid y) \cdot (B_k y + b_k)
\end{aligned}
$$

  further simplification, such as assuming $B_k = I$

Experiments

- TIMIT Corpus

- Four mode:
  - **No Compensation**: Acoustic models trained with full-bandwidth data and tested with band-limited data.
  - ***Model Adaptation****: Full-bandwidth acoustic models* adapted with data from the band-limited condition*.*
  - **Matched Models***: Acoustic models trained and tested with* band-limited data.
  - **CMN**: Models trained with CMN full-bandwidth data and tested with CMN limited-bandwidth data.

MODE	DISTORTION	%CORR	%ACC	DISTORTION	%CORR	%ACC
No Compensation	**Full-Bandwidth**	75.40	71.18			
CMN		75.71	71.61			
No Compensation	**LP6kHz**	64.32	58.30	**BP300-3400 Hz**	41.13	32.67
Model Adaptation		75.46	70.85		70.63	64.90
Matched Models		75.45	71.03		71.86	65.73
CMN		74.30	69.95		60.91	54.71
No Compensation	**LP4kHz**	55.93	44.67	**STC-TIMIT**	30.98	21.23
Model Adaptation		73.57	68.64		62.63	58.26
Matched Models		74.73	69.33		69.10	61.80
CMN		68.00	62.28		51.59	46.98
No Compensation	**LP2kHz**	30.45	26.10	**NTIMIT**	36.15	26.27
Model Adaptation		63.48	57.96		55.96	50.71
Matched Models		68.67	61.57		62.45	53.76
CMN		51.70	45.63		39.62	34.05

MODE	DISTORTION	%CORR	%ACC
No Compensation	**Full-Bandwidth**	75.40	71.18
CMN		75.71	71.61
No Compensation	**BP300-3400 Hz**	41.13	32.67
Model Adapt		70.63	64.90
Matched Models		71.86	65.73
CMN		60.91	54.71
Feature Compensation		70.62	64.79
CMN + Feature Comp.		70.12	64.31
M. Adapt + Feature Comp.		70.66	65.14
Matched M. + Feature Comp.		73.05	66.87
No Compensation	**STC-TIMIT**	30.98	21.23
Model Adapt		62.63	58.26
Matched Models		69.10	61.80
CMN		51.59	46.98
Feature Compensation		64.67	58.79
CMN + Feature Comp.		64.80	58.66
M. Adapt + Feature Comp.		66.25	60.12
Matched M. + Feature Comp.		71.32	63.96