

Bandwidth Mismatch Compensation for Robust Speech Recognition

¹Yuan-Fu Liao, ¹Jeng-Shien Lin and ²Wei-ho Tsai

¹Department of Electronic Engineering & Institute of Computer and Communication,
National Taipei University of Technology, 1, Sec. 3, Chung-Hsiao E. Rd. Taipei 106, Taiwan

²Institute of Information Science, Academia Sinica, Taiwan

¹yfliao@ntut.edu.tw, ²wesley@iis.sinica.edu.tw

Abstract

In this paper, an iterative bandwidth mismatch compensation (BMC) algorithm is proposed to alleviate the need of multiple pre-trained models for recognizing different bandwidth speech. The BMC uses the concept of the bandwidth extension as similar as in the speech enhancement approaches. However, it aims at directly improving the recognition accuracy instead of speech intelligence or quality and utilizes only recognizer's hidden Markov models (HMMs) for both bandwidth mismatch compensation and recognition. The BMC first detects the bandwidth of the input speech signal based on a divergence measurement. The HMM/Gaussian mixture model (GMM)-based method is then used to iteratively segment the input speech utterance and compensates the speech features. Experiments on serious bandwidth mismatched conditions, i.e., training on 8 kHz and testing on 4 kHz or 5.5 kHz bandwidth database have verified the effectiveness of the proposed approach.

1. Introduction

In a real-world application, the mismatch between training and test conditions often results in significant degradation on the performance of an automatic speech recognition (ASR) system. The mismatch may be due to the variations in speaker's characteristics, speaking style, transducer response, channel effect, background noise, and, especially, the difference in speech signal bandwidth.

Bandwidth mismatch between the training and test condition is often fatal due to inconsistent speech features even using the same sampling rate. For example, a HMM built from wideband (8 kHz) speech database is usually not feasible for recognizing up-sampled narrowband (4 kHz) speech.

However, it is not desired to keep multiple speech models for every possible bandwidth situations or to convert the information rich wideband speech back to the narrowband one. On the contrary, the high resolution wideband speech models with some form of compensation are preferred to reduce the mismatch and to further improve the recognition performance.

In order to utilize the wideband speech model, the bandwidth of the narrowband speech has to be extended. Many speech enhancement approaches for bandwidth conversion have been proposed including the statistical approach [1], the vector quantization (VQ) [2] and GMM mapping [3] and the HMM-based [4] methods.

The main body of the statistical approach is a pre-trained statistical recovery function which predicts the highband spectrum based on the narrowband spectrum. The idea behind

VQ for bandwidth extension is to create a codebook of corresponding wideband and narrowband speech features. The GMM with joint density estimation provides smooth spectral transformation. Moreover, for the codebook search, additional information from adjacent signal frames can be explored by utilizing the HMMs.

However, those speech enhancement methods all aim at improving the speech intelligence or quality but not at the recognition accuracy. Beside, additional mapping functions are usually required and have to be trained for every bandwidth mismatch conditions in advance.

In this paper, an iterative BMC algorithm based on the idea of directly improving the speech recognition accuracy is proposed. The GMM-based bandwidth extension method and the continuous large vocabulary speech recognizer (CLVSR) are integrated for recognizing different bandwidth speech signals. The key concepts are stated as follows:

- utilize the recognizer's HMMs for exploring the temporal structure information and use the state-dependent GMMs for precise bandwidth mismatch compensation
- use the parameters of the HMM state observation functions to construct a sequence of state-dependent GMMs
- use the discrete cosine transform (DCT) to transform the compensated filterbank features into MFCCs to partially absorb the high order compensation error/noise
- alternately segment the input narrowband speech in the MFCC domain and compensate it in the filterbank domain until converged

The remainder of this paper is organized as follows. In Section 2, the proposed BMC method is presented in detail. Experimental results showing the effectiveness of the proposed method are discussed in Section 3. Finally we summarize our findings in Section 4.

2. Proposed BMC approach

The proposed BMC method is an iterative algorithm which operates alternately in the filterbank and MFCC domain. The major function of the BMC includes the divergence-based missing filterbank feature vector components detection, the modified Viterbi search and the HMM/GMM bandwidth mismatch compensation. The functions and the detail procedures of the BMC algorithm are described in the following subsections.

2.1. Missing filterbank feature components detection

Although, some speech/audio file formats may have a header section which contains the sampling rate information, it is

still necessary to check whether there are any additional missing frequency components after upsampling the input utterance. A divergence-based reliability measure is computed in the filterbank domain and is compared with a pre-defined threshold R_H to detect the missing feature vector components. This zero-one decision will be used to guide the modified Viterbi search or the HMM/GMM bandwidth mismatch compensation method.

First, the symmetric divergence (or called Jeffrey's distance) [5] is used to measure the probabilistic distance between the distribution of the filterbank feature vector components of a test utterance and the distribution of the corresponding HMMs of the recognizer. The divergence of a distribution p with respect to another distribution q is defined as:

$$D(p \parallel q) = \int [p(x) - q(x)] \log \left(\frac{p(x)}{q(x)} \right) dx \quad (1)$$

The divergence could be greater than or equal to zero, and equals zero only when the two distributions are identical. If there are any missing filterbank feature vector components in a test utterance, the feature vector distribution is different and the divergence value will be large.

In the case of multivariate Gaussian distribution, the divergence measurement between two normal distributions $p = \mathbb{N}(\mathbf{u}_p, \boldsymbol{\Sigma}_p)$ and $q = \mathbb{N}(\mathbf{u}_q, \boldsymbol{\Sigma}_q)$, becomes [5]

$$D(p \parallel q) = \frac{1}{2} \left\{ \frac{(\mathbf{u}_q - \mathbf{u}_p)^T (\boldsymbol{\Sigma}_p^{-1} + \boldsymbol{\Sigma}_q^{-1}) (\mathbf{u}_q - \mathbf{u}_p)}{+tr(\boldsymbol{\Sigma}_p^{-1} \boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_q^{-1} \boldsymbol{\Sigma}_p - 2 \cdot \mathbf{I})} \right\} \quad (2)$$

The divergence measure is further converted into a reliability measure by embedding it into a smoothed zero-one function, e.g., the *sigmoid* function. The reliability measure is finally defined as follows:

$$R(D) = \frac{1}{1 + \exp(-\alpha D + \beta)}, \quad (3)$$

where α and β are the scaling and bias parameters of the *sigmoid* function. Thus, we expect to detect whether there are any missing filterbank feature vector components by Equation 3 in term of input speech features and the pre-trained speech models.

Furthermore, if the speech recognizer utilizes only diagonal covariance matrices, scalar forms of Equation 2 and 3 can be used to calculate the reliability measures make the detection decision for each feature vector component, i.e.,

$$D_d(p_d \parallel q_d) = \frac{1}{2} \left\{ \frac{(u_{q_d} - u_{p_d})^2 \left(\frac{1}{\sigma_{p_d}} + \frac{1}{\sigma_{q_d}} \right)}{+ \left(\frac{\sigma_{q_d}}{\sigma_{p_d}} + \frac{\sigma_{p_d}}{\sigma_{q_d}} - 2 \right)} \right\}, \quad (4)$$

$$R_d(D_d) = \frac{1}{1 + \exp(-\alpha D_d + \beta)}, \quad (5)$$

for $d = 1, \dots, M$, where M is the dimension of the filterbank feature vectors and $u_{p_d}, u_{q_d}, \sigma_{p_d}, \sigma_{q_d}$ are,

respectively, the means and variances of the d -th feature vector component.

2.2. Modified Viterbi search

After the missing filterbank feature vector components are detected, the likelihood calculation in the HMM Viterbi search procedure is modified to ignore those missing feature vector components [6]:

$$L\{\mathbf{X}, \mathbf{S} | \boldsymbol{\Lambda}\} = p(s_0) \prod_{t=1}^T p(s_t | s_{t-1}) \prod_{k=1}^K c_{s_t, k} \prod_{d, R_d < R_H} \mathbb{N}(x_{t,d} | u_{s_t, k, d}, \sigma_{s_t, k, d}) \quad (6)$$

where $\mathbf{X} = \{x_1, \dots, x_T\}$ is the input speech feature vector sequence, $\mathbf{S} = \{s_1, \dots, s_T\}$ is the state sequence, $\boldsymbol{\Lambda} = \{\mathbf{c}, \mathbf{u}, \boldsymbol{\sigma}\}$ is the parameters of the HMMs.

By this way, it is possible to use only one wideband HMM model to recognize different bandwidth speech signals in the filterbank domain. It is worth noting that by using Equation 6, the bad effect of the missing/inconsistent features could be alleviated.

2.3. HMM/GMM bandwidth mismatch compensation

To precisely compensate the bandwidth mismatch, the benefits of the HMMs and the GMMs are integrated. The HMMs are used to explore the temporal structure of the speech signal and the state-dependent GMMs are utilized to smoothly compensate the missing feature vector components.

Since the HMMs use the same mixture of Gaussian functions as the GMMs, the parameters of the HMMs could be reused to build the state-dependent GMMs through the inverse DCT (IDCT) transformation. For correct transformation between the filterbank and MFCC domain, extra high order MFCCs of the wideband speech are stored in the HMMs during the training procedure.

Given a segmentation/state sequence, the corresponding HMM parameters are therefore transformed by the IDCT to generate a sequence of state-dependent GMMs, $GMM(\bullet | s_t)$, in the filterbank domain. The detected missing filterbank feature vector components y_t for the t -th frame are then estimated/recovered from their corresponding narrowband filterbank feature vector components x_t according to the given state sequence and the re-constructed state-dependent GMMs [3] as follows:

$$\hat{y}_t = \frac{\sum_{k=1}^K c_{s_t, k} f_X(x_t | u_{s_t, k}^x, \sigma_{s_t, k}^x) \cdot u_{s_t, k}^y}{\sum_{k=1}^K c_{s_t, k} f_X(x_t | u_{s_t, k}^x, \sigma_{s_t, k}^x)} \quad (7)$$

The estimated wideband filterbank features $z_t = (x_t, y_t)$ are further converted into MFCC features to partially absorb the high order compensation error/noise by the DCT.

Moreover, it is possible to iteratively feed the compensated MFCCs back to the HMM recognizer for better

segmentation, and use the new segmentation for more precise filterbank features compensation. This iteration concept forms the basis of the proposed BMC algorithm.

2.4. The iterative BMC algorithm

The proposed BMC algorithm has two major phases, the initialization and the iteration procedures. The initialization procedure is used to absorb the defects of the missing filterbank feature vector components. The iteration procedure alternately re-compensates and re-segments the input speech for further improve the recognition performance. The detail algorithm iterates in the variable (l) is summarized below:

The BMC Algorithm

Initialization phase:

Step 1: Missing filterbank feature vector components detection

- 1) Compute the divergence measure D_d for each filterbank feature dimension d between the distribution p_d of the input filterbank feature $\mathbf{X}_{fb}^{(0)}$ and the distribution q_d of the corresponding pre-trained wideband HMMs Λ_{fb} in filterbank domain
- 2) Convert the divergence measure into the reliability measure $R_d(D_d)$ and compare it with the R_H

Step 2: Initial HMM segmentation

- 1) Find the first segmentation of the input speech $\hat{\mathbf{S}}^{(1)}$ by utilizing the modified Viterbi search and Λ_{fb}

$$\hat{\mathbf{S}}^{(1)} = \arg \max_{\mathbf{s}} L \left\{ \mathbf{X}_{fb}^{(0)}, \mathbf{S} \mid \Lambda_{fb} \right\} \quad (8)$$

Step 3: Initial HMM/GMM compensation

- 1) Recover the first version of the wideband filterbank feature $\hat{\mathbf{X}}_{fb}^{(1)}$ from its narrowband version $\mathbf{X}_{fb}^{(0)}$ according to the given segmentation $\hat{\mathbf{S}}^{(1)}$ and Λ_{fb}
- 2) Transform the compensated filterbank features $\hat{\mathbf{X}}_{fb}^{(1)}$ into the MFCC features $\hat{\mathbf{X}}_{mfcc}^{(1)}$ by the DCT

Iteration phase:

Step 4: HMM re-segmentation in MFCC domain

- 1) Use the recovered wideband MFCC features $\hat{\mathbf{X}}_{mfcc}^{(l)}$ and the HMMs Λ_{mfcc} to find a new segmentation $\hat{\mathbf{S}}^{(l+1)}$ of the input speech

Step 5: HMM/GMM re-compensation in filterbank domain

- 1) Use the given segmentation $\hat{\mathbf{S}}^{(l+1)}$ and Λ_{fb} to estimate a new version of the wideband filterbank features $\hat{\mathbf{X}}_{fb}^{(l+1)}$
- 2) Transform the compensated filterbank features $\hat{\mathbf{X}}_{fb}^{(l+1)}$ into the MFCC features $\hat{\mathbf{X}}_{mfcc}^{(l+1)}$ by the DCT

Step 6: go to Step 4 until converged.

3. Experiments

3.1. Databases

To examine the proposed method, a wideband microphone database was used. The speech signals were received and digitally recorded with a SoundBlaster card. A sampling rate of 16 kHz was used.

The database was divided into two subsets, nine-tenth of the database (referred to as TCC300-Training) was used for training the speech models; the remaining one-tenth of the database (referred to as TCC300-Test) was used for test. There are in total 274/29 speakers, 24,742/2,595 utterances, 300,856/31,411 syllables for training/test, respectively.

Furthermore, the test set, TCC300-Test, was down-sampled to 8 kHz or 11.025 kHz, separately, to limit the speech bandwidth to 4 kHz and 5.5 kHz, respectively. These two narrowband test sets were referred as TCC300-Test-4k and TCC300-Test-5.5k.

3.2. HMM recognizer under matched cases

Since we are interested in the training-test mismatch compensation problem due to bandwidth mismatch, a series of free-syllable decoding experiments were evaluated. In all experiments, continuous density HMM with left-to-right topologies and Gamma duration models were used.

The recognizer was gender-dependent with 100 right-context-dependent (RCD) *initials* and 40 context independent (CI) *finals* for each gender. The numbers of states and mixtures were empirically set to 3 and 5 states, each with maximum 32 mixtures, for *initial* and *final* HMMs, respectively. In addition, one single-state silence model with 64 mixtures was used. A 38-dimensional feature vector including 12 MFCC, 12 Δ -MFCC, 12 Δ^2 -MFCC, 1 Δ -log-energy and 1 Δ^2 -log-energy was used. The filterbank features were also evaluated for performance comparison. A 49-dimensional feature vector including 24 filterbank log-energies, 24 Δ -filterbank log-energies and 1 Δ -log-energy were used.

The syllable recognition results under matched cases, i.e., training and test both on the same 8 kHz, 5.5 kHz or 4 kHz bandwidth database were shown in Table 1. The best results while using MFCC features and the cepstral mean normalization (CMN) are 70.6%, 70.3% and 67.7% for 8 kHz, 5.5 kHz and 4 kHz bandwidth situation, respectively. It is worth noting that these results show the recognition rates while multiple HMMs are available for different bandwidth speech signals.

3.3. BMC algorithm under bandwidth mismatch cases

First, the backing-off method using the modified Viterbi search (see Equation 6) was tested in the filterbank domain. The recognition results are 56.9% and 59.5% (see Table 2 and 3) for the 4 kHz and 5.5 kHz bandwidth cases, respectively. The backing-off approach was also evaluated in MFCC domain by ignoring the missing features in DCT and used CMN to remove the channel bias. The results are 41.5% and 62.1% (see Table 2 and 3) for the 4 kHz and 5.5 kHz bandwidth cases, respectively. Both the filterbank and MFCCs results were much lower than the matched cases.

The BMC method was then initialized using the segmentation generated from the backing-off method. To absorb the high order compensation error/noise, the DCT was applied to transform the filterbank features into MFCCs. The CMN was also used to remove the channel bias. The results listed in Table 2 and 3 showed that the recognition rates were 56.5% and 59.1% using filterbank features and were boosted to 64.5% and 67.7% while using MFCCs for 4 kHz and 5.5 kHz bandwidth speech, respectively.

The BMC algorithms were then alternately iterated in the filterbank and MFCC domain. After 3 iterations, the recognition rates were raised to 64.7% and 67.8%, respectively. These results are higher than the filterbank and MFCC domain backing-off methods and are close to the matched cases. Finally, Figure 1 shows a typical compensated wideband filterbank features and its corresponding original wideband filterbank features. These evidences verified the effectiveness of the proposed BMC method.

4. Conclusions

We have proposed an iterative BMC algorithm to compensate the training-test bandwidth mismatch for robust ASR. Experiments on training on 8 kHz and test on 4 kHz or 5.5 kHz bandwidth speech databases, have verified that:

- It is feasible to use only one wideband model to recognize different bandwidth speech with little performance loss
- It is possible to recover/compensate the wideband speech features from its narrowband version
- DCT is capable of absorbing compensation error/noise

One area of further research is to study the bandwidth mismatch compensation under background noise or different channel conditions. The other one is to integrate other mismatch compensation/adaptation algorithm with BMC to further improve the recognition performance.

5. Acknowledgements

This work was supported in part by National Science Council (NSC), Taiwan, under contract NSC-91-2219-E-027-004 and in part by Ministry of Education (MOE) under contract EX-91-E-FA06-4-4. The authors also want to thank the Association for Computational Linguistics and Chinese Language Processing (ROCLING), Taiwan for supporting the TCC300 database.

6. References

- [1] Y.M. Cheng, D. O'Shaughnessy and P. Mermelstein, "Statistical Recovery of Wideband Speech from Narrowband Speech". *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 544-548, October 1994.
- [2] N. Enbom and W.B. Kleijn. "Bandwidth Expansion of Speech Based on Vector Quantization of the Mel Frequency Cepstral Coefficients". *IEEE Workshop on Speech Coding*, Porvoo, Finland, 1999.
- [3] K.-Y. Park and H.S. Kim. "Narrowband to Wideband Conversion of Speech using GMM-based Transformation". *Proc. ICASSP*, Istanbul, June 2000.

- [4] P. Jax and P. Vary, "Wideband Extension of Telephone Speech Using a Hidden Markov Model", *IEEE Workshop on Speech Coding*, pp. 133-135, Delavan, Wisconsin, September 2000.
- [5] P. A. Devijver and J. Kittler, "Pattern Recognition – A Statistical Approach," *Prentice-Hall International, London*, 1982.
- [6] J. Barker, M. Cooke, L. Josifovski and P. Green, "Soft Decisions in Missing Data techniques for Robust Automatic Speech Recognition," *ICSLP 2000*, Beijing.

Table 1: Recognition rates (%) for training and test both on the matched 8, 5.5 or 4 kHz bandwidth situation.

Bandwidth	Filterbank	MFCC+CMN
8 kHz	63.1	70.6
5.5 kHz	62.8	70.3
4 kHz	60.4	67.7

Table 2: Recognition rates (%) for training on 8 kHz and test on 4 kHz bandwidth situation.

	Filterbank	MFCC+CMN
Backing-off	56.9	41.5
BMC: Initialization	56.5	64.5
: 1 st Iteration	-	64.6
: 2 nd Iteration	-	64.6
: 3 rd Iteration	-	64.7

Table 3: Recognition rates (%) for training on 8 kHz and test on 5.5 kHz bandwidth situation.

	Filterbank	MFCC+CMN
Backing-off	59.5	62.1
BMC: Initialization	59.1	67.7
: 1 st Iteration	-	67.7
: 2 nd Iteration	-	67.8
: 3 rd Iteration	-	67.8

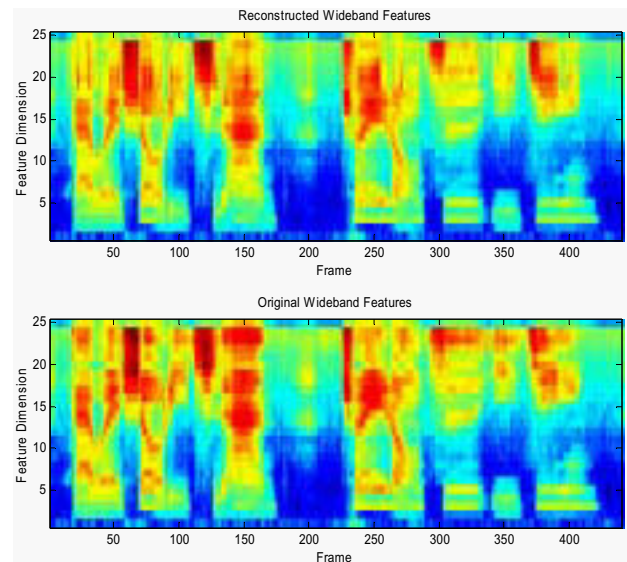


Figure 1: A typical recovered/compensated wideband filterbank features (top) and its corresponding original wideband filterbank features (bottom). In the top picture, the last seven filterbank feature vector components are all estimated from its narrowband (4 kHz bandwidth) version.