MFCC COMPENSATION FOR IMPROVED RECOGNITION OF FILTERED AND BAND-LIMITED SPEECH

Nicolás Morales[1,2], John H. L. Hansen[1] and Doroteo T. Toledano[2]

[1]Robust Speech Processing Group; Center for Spoken Language Research, Univ. of Colorado, Boulder USA
[2]HCTLab-Escuela Politécnica Superior. Univ. Autónoma de Madrid, SPAIN
e-mail:{nicolas.morales, doroteo.torre}@uam.es, john.hansen@colorado.edu

ABSTRACT

This paper addresses the problem of bandwidth expansion for the purpose of robust speech recognition. We show that an HMM-based ASR engine trained with full spectrum range data (0-8kHz) can successfully perform speech recognition tasks over band-filtered test data compensated by means of a series of simple MFCC parameter corrector functions. The problem is important when ASR is employed for audio streams of unknown frequency bandwidth common in spoken document retrieval. Evaluation is based on recognition rates. Accuracy varies depending on the width and spectral regions eliminated, but the system shows great advantages over the use of uncompensated filtered test data. The theoretical maximum recognition rates using corrector functions over filtered test data are very close to the base rate (unfiltered data) even when the greatest part of the spectrum of the original data is suppressed. These rates are even better than those obtained in the matched train/test HMMs with filtered data.

1. INTRODUCTION

The problem of frequency bandwidth extension is largely treated in the area of narrow-band speech enhancement. Many studies are motivated by telephone companies interested in the reconstruction of wideband speech (with frequencies ranging 0.05-7kHz) from a limited band-width transmission (0.3-3.4kHz). Low-band and high-band signal representations are inferred from the transmitted region, assuming a high degree of correlation, and are added to the input signal to reconstruct a wideband signal [1,2]. Similar studies exist for in-vehicle communications and other noisy environments [3,4]. Some of these studies have considered performing experiments using reconstructed speech with an extended frequency bandwidth; however, there has been limited literature specifically addressing the issue of bandwidth extension to improve ASR rates.

There are different situations where an ASR system trained with full-band data might be required to perform recognition with band-filtered unknown data. In this case, recognition performance is severely degraded.

Portable systems such as PDA's or mobile phones are likely to receive different bandwidth input signals (e.g., downloading media content from the Internet). Small footprint ASR engines are most suitable here, due to memory limitations. Thus, training multiple HMM sets at different bandwidths is not desirable.

Spoken document retrieval is another field that can benefit from our approach. The National Gallery of the Spoken Word contains historical speech records from the past 100 years, sampled at different rates [5]. Broadcast news recordings may contain audio with different bandwidths, for example when the news anchor talks with a field correspondent [6]. One major advantage of our approach is its high speed compared to training new models or using an HMM adaptation method. This allows a quick solution for rapidly changing environments.

Compensation of filtered data might also be useful in microphone mismatch in speaker identification [7], on-board car ASR [8] and extraction of the gist of activity in air traffic control [9] (where small airplanes may transmit with limited communication bandwidths).

Our problem is different from that of bandwidth extension for telephone communications in that there is no need for voice reconstruction. Thus, the cost and complexity needed to generate naturally sounding speech is avoided and the focus shifts to only spectral content. For the same reason, we are able to treat the problem directly in the MFCC domain, which considerably reduces computational costs.

2. MODELING THE EFFECT OF FILTERING IN THE MFCC DOMAIN

Our objective is to compensate restricted frequency input data for use in a generic ASR system without strong degradation in recognition rates. To achieve this, corrector functions are applied over the parameterized realization of

the input data. The front-end employs pre-emphasis filtering (α=0.97), using 25ms Hamming analysis windows with a 10ms window shift. The frame sequence is processed by using a bank of 26 triangular filters uniformly distributed in the Mel-Frequency scale along the region 0-8kHz. Finally, 13 MFCCs are computed (including C0) as well as their first and second derivatives. The use of a front-end based on a filter bank allows us to easily treat the problem in the MFCC domain.

Let us denote the output of each of the 26 Mel-Frequency filter bank channels as $fbank_j$ ($j = 1,...,26$). The MFCCs are computed from these values as [10]:

$$C_i = \sqrt{\frac{2}{26}} \sum_{j=1}^{26} \log(fbank_j) \cos\left(\frac{\pi i}{26}(j-0.5)\right) = \\ = \sum_{j=1}^{26} \log(fbank_j) A_{ij}, \quad i = 0,\ldots,12 \tag{1}$$

where A_{ij} represents the constant and the cosine function.

The effect of using filtered speech instead of the unfiltered version can be modeled as a multiplication of each of the values $fbank_j$ by a number, a_j, and the addition of an error term, e_j. This term models the unaccounted effects of filtering on all the speech processing and the errors due to the use of an FFT. The modified outputs of the filterbank, $\widehat{fbank}_j$ can be expressed as:

$$\widehat{fbank}_j = a_j fbank_j + e_j, \quad j = 1,\ldots,26 \tag{2}$$

Then, for the filtered speech, eq. (1) becomes:

$$\widehat{C}_i = \sum_{j=1}^{26} \left(\log\left(a_j \cdot fbank_j + e_j\right)\right) A_{ij} \tag{3}$$

In our experiments we use $a_j = 1$ for unfiltered and $a_j \to 0$ for filtered channels (this is valid for modeling both filtering and added frequency regions for oversampled data). Thus, the differences between the MFCCs of the unfiltered and filtered versions are:

$$C_i - \widehat{C}_i = \sum_{j=1}^{26} \Big[\log\left(fbank_j\right) - \\ - \log\left(a_j \cdot fbank_j + e_j\right)\Big] A_{ij} = \\ = \Bigg[\sum_{\substack{j=1 \\ j \notin F}}^{26} \Big[\log\left(fbank_j\right) - \log\left(fbank_j + e_j\right)\Big] + \\ + \sum_{\substack{j=1 \\ j \in F}}^{26} \Big[\log\left(fbank_j\right) - \log\left(a_j fbank_j + e_j\right)\Big]\Bigg] A_{ij} \tag{4}$$

where F represents the group of filtered channels. Assuming that $e_j \ll fbank_j$, the sum over the unfiltered channels in eq. (4) disappears. As $a_j \to 0$, for the filtered channels, $a_j \cdot fbank_j \ll e_j$. With this, eq. (4) becomes:

$$C_i - \widehat{C}_i = \sum_{\substack{j=1 \\ j \in F}}^{26} \Big[\log\left(fbank_j\right) - \log\left(e_j\right)\Big] A_{ij} \tag{5}$$

Thus, the differences in the MFCCs vary depending on the error values, e_j, as well as the energy of the original signal in each channel, $fbank_j$, and the filtered channels. We represent this dependence in terms of phonemes (namely a different correction is applied to each phoneme according to its spectral features in the filtered regions) while an unsupervised dependence based on classes could also be considered [11]. Intuitively, a low-pass filter will not seriously affect voiced phonemes (having most of their energy in low-frequency regions), but will more severely impact unvoiced phonemes such as fricatives.

3. COMPUTING COMPENSATION FUNCTIONS FROM FILTERED SPEECH

The most direct and precise way of computing compensation functions in the MFCC domain is comparing the MFCCs obtained from unfiltered and filtered speech for each phoneme. This way, both the dependence on the phonemes and the error term in eq. (5) are modeled. In our experiments, compensation functions are obtained by mapping unfiltered to filtered data using the TIMIT training partition. For each file, both the filtered and unfiltered versions are parameterized, thus generating a filtered frame for each frame of the unfiltered file. Data time-labels are available so it is possible to identify each frame window with its corresponding phoneme. This allows a different mapping for each phoneme and MFCC parameter (as suggested by eq. (5)). Compensation functions are computed as the 5th order polynomial fits of this mapping procedure. Fig.1 shows an example for phoneme /ae/, MFCC coefficient C4 and 4kHz low-pass filter. The compensation function is plotted too. Although eq. (5) shows that the compensation should be dependent on the energy distribution, for practical reasons it is desirable to have also a general phoneme-independent compensation computed on data from all the phonemes. This less precise compensation has the advantage of being applicable to all phonemes.

4. COMPUTING COMPENSATION FUNCTIONS FROM FILTERED FBANK CHANNELS

A less precise but much more efficient (in terms of computation) way of generating the compensator functions is ignoring the error term in eq. (5) and assuming that the filtering operation can be modeled as a multiplication of the outputs of the filter-bank channels

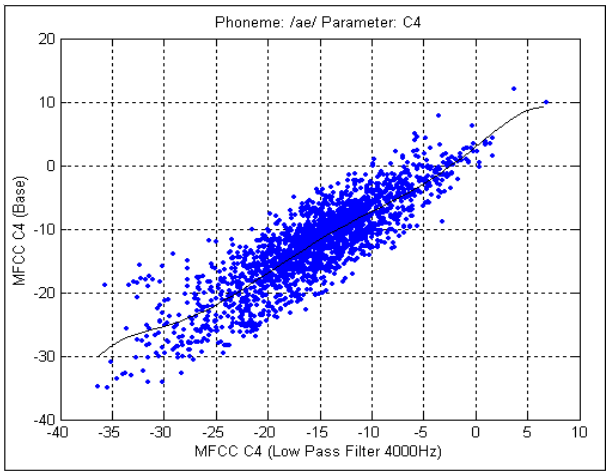

Figure 1: Polynomial fit for MFCC 4 of phoneme /ae/. Low-pass 4kHz filter.

(Mel-Frequency Energies, MFEs) by the filtering coefficients, a_j, dependent only on the filter applied. In order to avoid values near zero in the log calculation, a floor value is used for the argument of the logarithm. A block diagram of this strategy is shown in Fig.2.

As discussed in Sec.2, filtering the MFEs is an approximation to filtering the training sound files and finding the MFE coefficients of the filtered data. However this is much faster than filtering the sound files. From the MFE coefficients it is straightforward to obtain the MFCCs, and mapping is performed as in Sec.3.

5. IMPLEMENTATION

As will be shown, the compensation functions do model the transformation of the MFCCs very successfully. However, the main problem for ASR implementation is to know which phoneme compensation formula (which 5th order polynomial fit) to apply in each frame of the unknown input utterance. Ideally, different compensations should be applied for each phoneme, but before recognition there is no information on the phonemes and boundaries. We propose the block diagram system in Fig.3. First, a general compensation (based on data from all the phonemes and silences) is applied over the input MFCC representation. Phoneme-level speech recognition follows, generating several transcription candidates. Each candidate allows to create a different phoneme-specific compensation of the original. A word-level ASR is then applied over each of the compensated versions and the best is chosen as in a ROVER system [12].

6. RESULTS AND DISCUSSION

Evaluation is performed using HTK tools [10]. An HMM-based ASR engine is trained using the training partition of TIMIT. Fifty-one models (3 states with 15 Gaussian

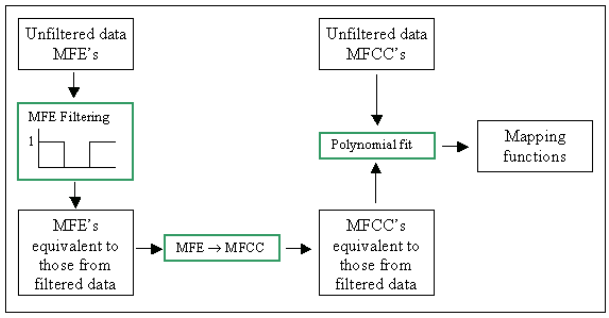

Figure 2: Compensator functions calculated filtering MFC coefficients.

mixtures each) are trained including short-pause and starting and ending silences. The front-end features are 13 MFCC coefficients with C0 and their respective first and second derivatives for a total of 39 parameters.

Table 1 is a summary of results for different bandwidth filters. Here, the base system refers to unfiltered full-band 8kHz data. Specific correction implies the use of time labels in the correction phase followed by an unsupervised ASR. General correction functions are those calculated with data from all phonemes. ROVER refers to the 2-stage correction scheme in Fig.3. For the case of 2-4kHz band-cut, a test has been run for specific correction formulas obtained by filtering the MFEs of the training data (Sec.4).

For each evaluated filter, uncompensated filtered data represents the starting accuracy rate and the base system accuracy rate is the maximum attainable.

Specific correction results are very close to those of the base system even for the case of 2kHz low-pass filter (i.e. trained with 8kHz bandwidth data and tested with 2kHz bandwidth data). This means that corrector functions are capable of modeling the shift in the MFCCs due to data filtering. For the cases of low-pass 4kHz and 2kHz, we show the accuracy rate obtained by model adaptation using an MLLR+MAP schema [10] and HMM training with filtered data. Specific correction performs better than either of the other two, proving the ability to recover some of the information lost by filtering. However, the application of phoneme-specific correction is not straightforward in real conditions.

General correction is equivalent to applying a general rotation to the MFCC space. This is applicable in real situations but its performance strongly depends on the frequency bands eliminated, as shown in Table 1.

The 2-stage ROVER correction schema proves successful for low-pass 6kHz and 4kHz filters (with rates similar to those of model adaptation and filtered data HMM training). The success of this method is highly correlated with a successful general correction.

Finally, evaluation of the corrector functions computed by filtering MFEs (for 2-4kHz band-cut filter), shows a slightly reduced recognition rate (compared to

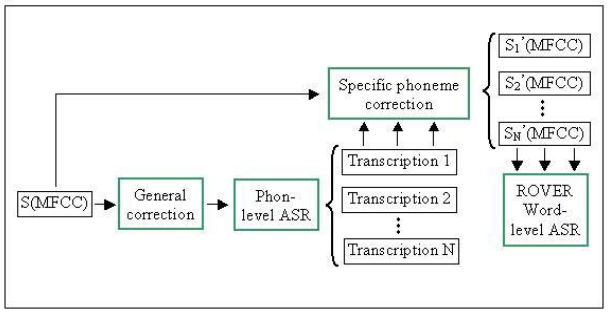

Figure 3: Real case implementation.

specific compensation functions calculated from filtered speech files) due to ignoring the error term in eq. (3).

7. CONCLUSIONS AND FUTURE WORK

We have shown the potential importance of narrow band data enhancement for robust speech recognition. Corrector functions mapping filtered realizations to their corresponding unfiltered versions show very promising results even in the case of large band filtering. Specific correction results proved superior to those for model adaptation or training of models with filtered data. Throughout this work we focused on adapting an MFCC based front-end, keeping the base models untouched.

More work is needed for use in real situations, especially for large band filtering. However, in situations such as ASR for enrollment of unknown audio in spoken document retrieval the ability to overcome varying frequency bandwidths can significantly improve text transcription results. The accuracy rate is still far from the theoretical maximum (phoneme-specific correction) due to the difficulty of finding phoneme locations and boundaries. The first stage (general correction) in the ROVER scheme has to be improved so that the effective phoneme-specific corrections can be adequately applied.

The main advantages of our approach compared to other solutions such as multiple HMM sets trained at different sampling rates or model adaptation are the potential increase in accuracy rate, the compactness of the system (i.e. memory requirements) and the short time needed to calculate corrector functions. The computation of corrector functions from filtered MFEs optimizes the process and makes it possible to obtain specific corrector functions in real time.

Test Set	HMM train set	Correction	Percent Correct	Percent Accuracy
Base	Base	None	77.18	75.31
Low-Pass 6k	Base	Specific	77.14	75.28
	Base	None	56.55	52.27
	Base	General	75.40	72.98
	Base	ROVER	76.62	74.57
Low-Pass 4k	Base	Specific	76.55	73.01
	Base	None	31.67	23.20
	Base	General	66.56	62.00
	Base	ROVER	71.36	67.61
	Base	Model Adapt.	73.55	71.08
	LP4000	None	72.92	70.10
Band-Cut 2-4k	Base	Specific	75.15	73.55
	Base	None	1.84	-0.06
	Base	General	60.63	57.74
	Base	ROVER	64.57	62.98
	Base	Specific (filt. cepstra)	71.98	68.28
Low-Pass 2k	Base	Specific	72.36	69.56
	Base	None	2.11	-0.56
	Base	General	26.24	18.10
	Base	ROVER	32.78	26.37
	Base	Model Adapt.	62.17	58.56
	LP2000	None	62.87	56.66

Table 1: Baseline results and results with different degrees of filtering and compensation strategies.

8. REFERENCES

[1] G. Miet, A. Gerrits and J.C. Valiere, "Low-band extension of telephone-band speech", *Proc. ICASSP'00* **(3)**. 1851 – 54.

[2] S. Chennoukh, A. Gerrits, G. Miet and R. Sluijter, "Speech enhancement via frequency bandwidth extension using line spectral frequencies", in *Proc. ICASSP'01* **(1)**. 665 – 668.

[3] C. Aimin, S. Vaseghi and P. McCourt, "State based sub-band LP Wiener filters for speech enhancement in car environments", in *Proc. ICASSP'00* **(1)**. 213 – 216.

[4] H. Schnepp, T. Muller, J.-F. Luy and P. Russer, "The implementation of channel diversity in mobile software radio receivers", in *IEEE Microwave and Wireless Components Letters* **(13)**: 8. 323 – 325, Aug. 2003.

[5] http://www.ngsw.org/

[6] J.H.L. Hansen, B. Zhou, M. Akbacak, R. Sarikaya and B.L. Pellom, "Audio Stream Phrase Recognition for a National Gallery of the Spoken Word", *Proc. ICSLP'00*.

[7] D. Reynolds and X. Wang, "Adaptive transmitter optimization for blind and group-blind multiuser detection", in *IEEE Trans. Signal Processing* **(51)**: 3. 825 – 38, March 2003.

[8] H. Abut, J.H.L Hansen and K. Takeda (eds.), *DSP for In-Vehicle and Mobile Systems*, Kluwer/Springer-Verlag. 2005.

[9] L. Denenberg, H. Gish, M. Meteer, T. Miller, J.R. Rohlicek, W. Sadkin and M. Siu, "Gisting conversational speech in real time", in *Proc. ICASSP'93* **(2)**. 131 – 134.

[10] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, W. Valtchev and P Woodland, *The HTK Book (for HTK version 3.2.1),* December 2002.

[11] Y. M. Cheng, D. O'Shaughnessy and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech" *IEEE Trans. Speech and Audio Processing* (2): 4. 544-48, Oct. 1994.

[12] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)", *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding* . 347 - 354, 14-17 Dec. 1997.