UNSUPERVISED CLASS-BASED FEATURE COMPENSATION FOR TIME-VARIABLE BANDWIDTH-LIMITED SPEECH

Nicolás Morales[1,2], *Doroteo T. Toledano*[3], *John H. L. Hansen*[2], *Javier Garrido*[1] *and José Colás*[1]

[1]HCTLab, [3]ATVS Lab., Universidad Autónoma de Madrid, Spain.
[2]CRSS, University of Texas at Dallas, USA.
e-mail:{nicolas.morales, doroteo.torre, javier.garrido, jose.colas}@uam.es, john.hansen@utdallas.edu

ABSTRACT

This paper deals with the problem of speech recognition on band-limited speech. In our previous work we showed how a simple polynomial correction framework could be used for compensation of band-limited speech to minimize the mismatch using full-bandwidth acoustic models. This paper extends this approach to time-varying multiple-channel environments. The compensation framework is extended to perform automatic channel classification prior to compensation, thus allowing for unsupervised multi-channel compensation without the need for an explicit channel classifier. Performance is demonstrated on a wide range of channel bandwidth conditions. This extension makes our compensation approach potentially applicable in a much wider range of scenarios with only very limited performance degradation compared to the supervised approach.

1. INTRODUCTION

Robustness is one of the main challenges in speech recognition systems today. While systems perform successfully under ideal circumstances (typically those of a laboratory) recognition rates tend to decrease rapidly when the system conditions are changed. Mismatch between training and testing conditions is in general an unsolved problem and due to its practical interest it has at the present time great relevance in the field of speech recognition.

In this paper we study the mismatch created by a channel removing part of the available speech spectrum. It is well known that a system trained with full-bandwidth speech (here we consider full-bandwidth as the range of frequencies from 0 to 8 kHz) fails to perform speech recognition with band-limited speech material. The typical solution is training the recognizer with band-limited data that matches the test conditions. However, this requires an ad-hoc system trained for each particular channel condition, thus incurring in significant resources in data collection as well as retraining each time a system is designed.

Our approach performs speech feature compensation, which allows using a single acoustic model set under a variety of bandwidth limitations (typically caused by the transmission channel). The parametric representation (features) of band-limited speech is compensated using simple polynomial functions trained to map band-limited feature vectors to full-bandwidth feature vectors. The idea was inspired by classical bandwidth extension approaches which aim at the reconstruction of the complete spectrum to obtain more natural speech signals [1,2] − a complicated task because the human ear is very sensitive to artifacts in the bandwidth-extended speech. However, as we showed in previous work [3,4], when the purpose of bandwidth extension is speech recognition, the problem may be greatly simplified by reconstructing only the features (not the speech itself), as the amount of information to be reconstructed is highly reduced. Also, the reconstruction may be performed directly in the space of Mel Frequency Cepstrum Coefficients (MFCC), the parameterization of choice for most speech recognition systems, which allows for easy integration in off-the-shelf systems.

One of the limitations of our previous compensation framework was that it needed to be supervised (i.e. it was necessary to know in advance the type of bandwidth limitation). This paper builds on our previous work, which we extend here to allow for unsupervised compensation of speech coming from a range of different and possibly time-varying band limiting channels.

Our new approach can be used when the input bandwidth limitation is unknown, as is the case with spoken document retrieval [5] or broadcast news' transcription, where depending on the various recording or transmission technologies the speech may present different bandwidth limitations. It also allows for self-adaptation when the input channel varies over time, as can be seen in broadcast material, where full-bandwidth speech (typically when the anchor talks in the studio) may be mixed with limited speech from recordings, telephone, or field correspondants. Our new approach would also be useful on systems receiving speech from different sources such as PDAs or cell phones (incoming speech from the internet or different telephone operators), or environments such as the control tower in an airport receiving speech from different radio devices [6].

1-4244-0469-X/06/$20.00 ©2006 IEEE

2. MODEL OF BANDWIDTH-LIMITATION CHANNEL DISTORTION

In [3] we developed our model framework of the effect of channel bandwidth-limitation. Here we present the most important conclusions.

The MFCC representation of the full-bandwidth signal is typically expressed as:

$$x_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} \log(fbank_j) \cos\left(\frac{\pi i}{N}\left(j-0.5\right)\right) = \\ = \sum_{j=1}^{N} \log(fbank_j) \cdot A_{ij} \tag{1}$$

where x_i is the MFCC of order i and N is the number of channels in the filterbank. Using a similar expression for the MFCCs of the band-limited signal (y_i) and subtracting both terms we obtained:

$$x_i - y_i = \sum_{\substack{j=1 \\ j \in F^b}}^{N} \left[\log\left(fbank_j\right) - \log\left(e_j\right)\right] A_{ij} \tag{2}$$

where F^b is the group of filterbank channels affected by the distortion and e_j is an error term introduced by non-ideal filtering as well as finite-precision effects.

The following sections describe our strategy to "fill the gap" left in the MFCC by the channel distortion so that a system trained with full-bandwidth speech may perform similarly to recognizers trained with band-limited speech (matched recognizers).

3. GAUSSIAN CLASS-BASED LINEAR CORRECTION

In our previous work on supervised bandwidth extension [3,4], we demonstrated the convenience of partitioning the MFCC space of a particular bandwidth limitation by means of Gaussian mixtures, as suggested in [7]:

$$p\left(\boldsymbol{y}\right) = \sum_{k^b=1}^{K^b} p\left(\boldsymbol{y}\middle|k^b\right) \cdot P\left(k^b\right) = \\ = \sum_{k^b=1}^{K^b} N\left(\boldsymbol{y}, \mu_{k^b}, \Sigma_{k^b}\right) \cdot P\left(k^b\right) \tag{3}$$

where $\boldsymbol{y}$ are the MFCC vectors of the speech distorted (band-limited) by channel b. The MFCC space corresponding to speech with different bandwidths can straightforwardly be represented as:

$$p\left(\boldsymbol{y}\right) = \sum_{b=1}^{B} \sum_{k^b=1}^{K^b} N\left(y, \mu_{k^b}, \Sigma_{k^b}\right) \cdot P\left(k^b\right) = \\ = \sum_{\psi=1}^{\Psi} N\left(y, \mu_{\psi}, \Sigma_{\psi}\right) \cdot P\left(\psi\right) \tag{4}$$

We note $\Psi = \{\psi\} = K^1 \bigcup K^2 \bigcup \ldots \bigcup K^B$ as the super-set of Gaussian classes grouping together the sets of Gaussian mixtures K^b from different input distortions.

Assuming that within each class the full-bandwidth MFCC vectors $\boldsymbol{x}$ and their limited counterparts $\boldsymbol{y}$ are jointly Gaussian, the probability of the full-bandwidth vector $\boldsymbol{x}$ given mixture ψ and limited-bandwidth vector $\boldsymbol{y}$ is also Gaussian and its expectation is:

$$\boldsymbol{x}\left(\boldsymbol{y},\psi\right) = E\left\{\boldsymbol{x}\middle|\boldsymbol{y},\psi\right\} = \\ = \mu_{\boldsymbol{x}}^{\psi} + \Sigma_{\boldsymbol{xy}}^{\psi}\left(\Sigma_{\boldsymbol{y}}^{\psi}\right)^{-1}\left(\boldsymbol{y} - \mu_{\boldsymbol{y}}^{\psi}\right) = \boldsymbol{B}^{\psi}\boldsymbol{y} + b_0^{\psi} \tag{5}$$

The estimation of matrix $\boldsymbol{B}^{\psi}$ and the constant b_0^{ψ} is complex and given the error terms introduced in the filtering process, a precise computation does not necessarily mean better performance. For simplicity, we choose a diagonal matrix $\boldsymbol{B}^{\psi}$, an approximation that given the low degree of correlation between the MFCC parameters should not result in a substantial loss in performance. The correction applied to the MFCC of order i and time frame t is then:

$$x_i\left(t\right) \approx b_1^{\psi,i} \cdot y_i\left(t\right) + b_0^{\psi,i} \tag{6}$$

which is a first order polynomial correction of the band-limited MFCC value.

4. GAUSSIAN CLASSES AND CORRECTOR FUNCTIONS TRAINING

Classes from different band-limiting channels are trained separately, allowing for independent optimizations as well as to successively add new channel-dependent corrector classes as needed.

All training data from a given band-limiting distortion is initially pooled together in a single class and a top-down schema is followed as the number of Gaussian classes dividing the space is incremented one by one (the number of final classes needs to be found empirically. In this study 32 classes were trained for each distorting environment).

Once the classes dividing the distorted space are created, a different linear corrector function is trained for each Gaussian class by finding the coefficients for Eq. (6) that best represent the mismatch. For this task, we use stereo data (for each available training frame from a given channel limitation, there is a corresponding full-bandwidth frame). We obtain the mapping for each MFCC parameter from their limited bandwidth values y_i to their corresponding full-bandwidth values x_i. The coefficients that reflect this fit are then determined by linear regression.

5. UNSUPERVISED CLASS BASED CORRECTION

Gaussian classes trained with data from different channels are combined in a super set of classes, and for each input frame we compute the compensation corresponding to each

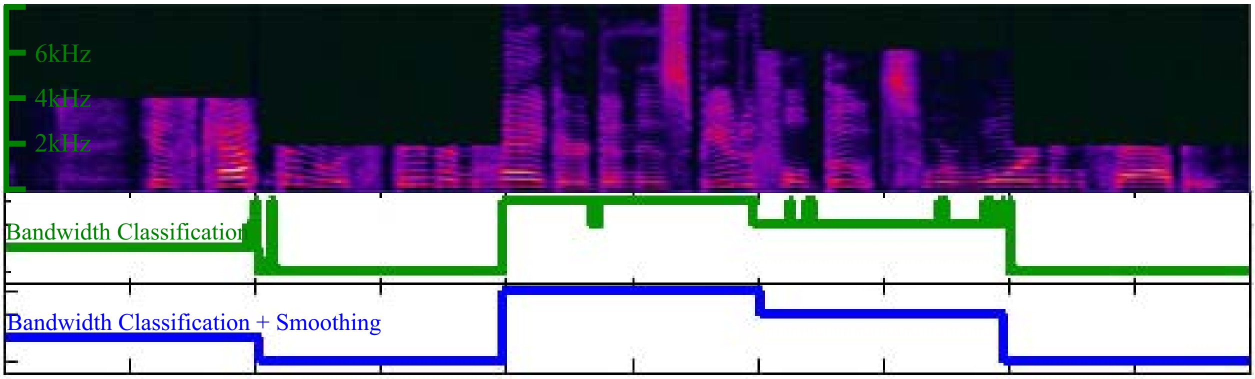

Figure 1: Automatic detection and smoothing the output of variable bandwidth limitation on file DR1_FAKS0_SI1573 of TIMIT. The figure shows that smoothing the output classification removes all punctual errors.

class in the super-set. The final compensation value is calculated as the following weighted average:

$$x_i(t)=\sum_{\psi=1}^{\Psi} p(\psi|\boldsymbol{y}(t))\cdot\left(b_0^{\psi,i}+b_1^{\psi,i}\cdot y_i(t)\right) \qquad (7)$$

where $p(\psi|\boldsymbol{y}(t))$ is the posterior probability of class ψ given the observed MFCC vector $\boldsymbol{y}$.

In general one can assume that the input channel will not vary continuously, but will remain constant for some amount of time (in broadcast speech recognition, for example, speech may come from different sources with different channel distortions, but information from different channels would typically last at least a few seconds). Thus, the bandwidth classification output may be refined by smoothing within windows of length N frames (Figure 1).

6. RESULTS AND DISCUSSION

Evaluation is performed using TIMIT. The features used are 13 MFCCs with deltas and double deltas. Different band-limiting channels are simulated by filtering the speech files with step-like low-pass filters with cut-off frequencies of 6kHz, 4kHz and 2kHz, respectively. The training partition of TIMIT is used for training full-bandwidth context independent phonetic HMM models. Fifty-one models with three emitting states are trained. Gaussian classes as well as the corrector functions' coefficients (Eq. (6)) are trained using a small set of the training partition.

In our experiments we evaluate the ability of our method to classify the distorting channel, as well as the phoneme recognition performance achieved with automatic classification and associated correction.

6.1. Bandwidth-limitation classification accuracy

A multi-channel correction system was created combining Gaussian classes trained with data from the low-pass filters considered (cut-off frequencies at 6kHz, 4kHz and 2kHz, respectively) as well as unfiltered data. For evaluation, the speech files in the test partition of TIMIT are passed through each of the considered channels separately and for each frame we evaluate the correctness of the most likely class (namely if the class belongs to the set trained with the actual bandwidth limitation applied).

Table 1 shows classification results for each of the frequency limitations considered. As can be seen, classification is less accurate for less restrictive low-pass filters. As most of the energy is contained in the lower parts of the spectrum, filtering the region 6000-8000 Hz results in a relatively small distortion and the confusability between full-bandwidth and 6kHz low-passed frames is relatively high. At the same time, and for the same reason, we expect misclassification between these two channels not to incur in very significant correction differences, thus resulting in only a small impact on phonetic recognition performance.

Classification ability was also evaluated when different filters were applied to different segments of the signal. Results were very similar to those reported in Table 1, for example for a low-pass filter with cut-off frequency 6kHz, classification accuracy is 92.73% (98.60 with a mode filter). Classification degradation is only observed in the 2 or 3 frames before and after the channel variation, which will be silent portions in many practical cases.

In Figure 1 we show the filtering channel classification across time for an utterance for which filtering varied every second. As can be seen, the misclassification regions are narrow and may be repaired by smoothing the classification output (we used 21 frame-long smoothing windows).

6.2. Phonetic recognition performance

Next, we evaluate the impact caused on phonetic recognition accuracy by misclassification of the distorting channel. Full-bandwidth acoustic models are used for recognition of limited bandwidth speech from different channels as well as full-bandwidth speech. Results are given for input speech corrected using both, supervised and unsupervised Gaussian-class correction, as well as Cepstral Mean Normalization (CMN), and no compensation. Results are also given for matched models, which is the typical (supervised) approach.

Table 2 shows that the impact introduced by passing from unsupervised to supervised recognition is very limited

Detected Chan →	**FB**	**LP6000**	**LP4000**	**LP2000**
Input Channel ↓				
FB	91.58 **(98.81)**	6.57 **(0.99)**	1.61 **(0.18)**	0.25 **(0.02)**
LP6000	5.07 **(0.18)**	92.85 **(99.43)**	1.94 **(0.38)**	0.14 **(0.01)**
LP4000	1.25 **(0.01)**	0.46 **(0.00)**	97.96 **(99.95)**	0.33 **(0.03)**
LP2000	0.19 **(0.00)**	0.03 **(0.00)**	0.02 **(0.00)**	99.76 **(100.00)**

Table 1: Input distortion classification rates (in %). Results in parentheses are obtained using bandwidth output smoothing.

thanks to the success in channel bandwidth classification. For example, for input data with a low-pass filter of 4kHz, the relative loss of unsupervised compensation compared to supervised compensation is only 0.5% and negligible if smoothing is employed. For all the bandwidth limitations considered, performance is clearly superior to that of the baseline (no compensation) or with CMN. Compared to an ad-hoc solution like matched models, performance is relatively close except for a low-pass filter with cut-off frequency 2kHz. This result is very encouraging since our unsupervised approach can seemingly deal with speech with 4, 6 and 8 kHz bandwidths with performances very close to that of the matched models solution and requiring only one set of acoustic models.

7. CONCLUSIONS AND FUTURE WORK

In this work we show how class-based polynomial correction for band-limited speech may be integrated in an unsupervised system for correction of speech from different band-limiting channels. The system determines the nature of the input channel and successfully employs the corrector functions matching the channel distortion for compensation of band-limited speech. Full-bandwidth speech (8kHz) was also considered for completeness, showing that this compensation scheme can be introduced without significant impact in systems working under a variety of conditions.

It was also shown that assuming that channel limitations remain stable for at least the duration of a few phones, unsupervised channel output classification may be smoothed so that speech recognition performance is almost identical to that of the supervised compensation system.

The success of our unsupervised approach highly increases the utility of our previous feature-based compensation of bandwidth limitation for speech recognition. Performance is similar to that of the best method known – training and using matched models. Furthermore, our unsupervised approach can also be used even under unknown and time-variable bandwidth limitations such as those encountered in spoken document retrieval or broadcast news' speech recognition.

In the future we will investigate how our framework scales to a larger number of filtering conditions and also how well it generalizes to more realistic environments such as telephone speech and broadcast news.

Input channel	**Compensation**	**Percent Correct**	**Phoneme Accuracy**
F-B	Unsupervised	75.22	70.79
F-B	Unsupervised (+smooth)	75.33	71.08
F-B	No Compensation	75.40	71.18
LP6000	Unsupervised	74.81	70.48
LP6000	Unsupervised (+smooth)	74.84	70.62
LP6000	Supervised	74.88	70.67
LP6000	Matched	75.45	71.03
LP6000	CMN	74.30	69.95
LP6000	No Compensation	64.32	58.30
LP4000	Unsupervised	72.31	66.81
LP4000	Unsupervised (+smooth)	72.32	67.10
LP4000	Supervised	72.32	67.11
LP4000	Matched	74.73	69.33
LP4000	CMN	68.00	62.28
LP4000	No Compensation	55.93	44.67
LP2000	Unsupervised	55.07	48.22
LP2000	Unsupervised (+smooth)	55.01	48.21
LP2000	Supervised	55.01	48.21
LP2000	Matched	68.67	61.57
LP2000	CMN	51.70	45.63
LP2000	No Compensation	30.45	26.10

Table 2: Phonetic classification results for different input channel bandwidth limitations and compensation approaches.

8. ACKNOWLEDGEMENTS

This work was supported partly by the Spanish Ministry of Education, RADC under contract SC-05-02/S27410, and the State of Texas under Project Emmitt.

9. REFERENCES

[1] Y.M. Cheng, D. O'Shaughnessy, P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech", Speech and Audio Processing, IEEE Trans. Speech and Audio, (2): 4. 544-548, Oct. 1994.

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

[3] N. Morales, J.H.L. Hansen and D.T. Toledano, "MFCC Compensation for Improved Recognition of Filtered and Band-Limited Speech", *Proc. ICASSP'05,* (1): 521-524.

[4] N. Morales, D.T. Toledano, J.H.L. Hansen, J. Colás and J. Garrido, "Statistical Class-Based MFCC Enhancement of Filtered and Band-Limited Speech for Robust ASR", Proc. InterSpeech'05, 2629-2632.

[5] J.H.L. Hansen, R. Huang, B. Zhou, M. Seadle, J.R. Deller Jr., A.R. Gurijala, M. Kurimo and P. Angkititrakul, "SpeechFind: Advances in Spoken Document Retrieval for a National Gallery of the Spoken Word", IEEE Trans. Speech and Audio, 13.5: 712-730.

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

[7] X. Huang, A. Acero and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall PTR, 2001.