



Narrowband to Wideband Feature Expansion for Robust Multilingual ASR

Dušan Macho

Center for Human Interaction Research, Motorola Labs, Schaumburg, USA

dusan.macho@motorola.com

Abstract

To build high quality wideband acoustic models for automatic speech recognition (ASR), a large amount of wideband speech training data is required. However, for a particular language, one may have available a lot of narrowband data, but only a limited amount of wideband data. This paper deals with such situation and proposes a narrowband to wideband expansion algorithm that expands the narrowband signal ASR features to wideband ASR features. The algorithm is tested in two practical situations comprising sufficient amount and insufficient amount of original wideband training data. Tests show that using a combination of wideband features and expanded features does not harm the ASR performance when having a sufficient amount of the original wideband data, and it improves the ASR performance significantly when only a limited amount of wideband data is originally available. In the presented multilingual tests, a unique expansion model is trained for four languages from the Speecon database. Availability of different amounts of wideband training data is considered, including the case when no wideband data is available. ASR experiments for each language confirm that the addition of expanded features to the wideband model training enhances the models and provides better results than using the limited amount of wideband data only. In all tests, the ETSI standard noise-robust front-end is used.

Index Terms: narrowband to wideband feature expansion, multilingual speech recognition, robust speech recognition

1. Introduction

There is an interest in building wideband acoustic models for ASR due to their potential of achieving a superior performance in comparison to narrowband models. However, collecting wideband speech training data is more difficult and costly than collecting narrowband speech data, mostly due to the existing telecommunication infrastructure and standards. Thus, when building ASR acoustic models for several languages, for a particular language there may be available only a small amount of wideband data or, in an extreme case, no wideband data, but the amount of narrowband data may be large. This paper proposes an approach that allows building wideband acoustic models using only a small amount or no wideband speech data and a large amount of narrowband data by performing a narrowband to wideband (NB2WB) expansion of narrowband ASR features. Throughout this paper, it is referred to an 8 kHz sampled signal as narrowband signal and to a 16 kHz sampled signal as wideband signal.

The idea of NB2WB expansion is not new. Several approaches have been used so far [1]-[5]. In these approaches, either narrowband signal or narrowband features are expanded by using the linear model of speech production [1], previously trained statistical model such as Gaussian mixture model (GMM) [2] or hidden Markov model (HMM) [3],[4], or more recently neural network model [5]. The advantage of

the HMM approach is its closeness to our target application, ASR, and a possibility of exploring the temporal alignment of narrowband signal with the wideband model. Thus, the HMM-based ASR feature expansion approach is used in this paper. Two practical situations that may occur when training ASR acoustic models are considered: sufficient and insufficient amount of wideband training data. We observe the effect of adding the NB2WB expanded data into the training data set in both of these situations. In the second part of experiments, the proposed NB2WB expansion is applied to several languages using a unique HMM expansion model. Additionally, a wideband ASR acoustic model is trained even for a language for which no wideband training data is available.

2. Proposed narrowband to wideband expansion

Narrowband signal contains no information at high frequencies ranging from 4 to 8 kHz. The objective of NB2WB expansion is, assuming availability of low frequency information from 0 to approximately 4 kHz, to recover the missing high-frequency information. The presented NB2WB expansion algorithm uses a wideband HMM acoustic model and performs Viterbi alignment to calculate the missing high-frequency filter-bank (FB) bands in the feature domain. In principle, it is similar to the approach mentioned in [3] or [4]. The algorithm consists of the following steps:

Step 1: Building expansion HMM (xHMM)

In this step, a phoneme xHMM is trained using all available wideband data. The training process is depicted at Figure 1 where "WB Data" denotes the available wideband data. Speech features should be in the log filter-bank domain because the missing high-frequency components are clearly localized in this domain. Note that in Section 3.2 modified filter-bank features are proposed in this work to achieve a better performance.

Step 2: State alignment of narrowband data with xHMM

In this step, the time sequence of feature vectors of narrowband signal $\text{seq}^{NB} = \{\mathbf{x}^{NB}(1), \mathbf{x}^{NB}(2), \dots, \mathbf{x}^{NB}(T)\}$ is aligned with the xHMM using the Viterbi alignment at state level. For simplicity, the time index of feature vector is dropped in the following formulas and a feature vector consisting of N elements $\mathbf{x}^{NB} = \{x_1^{NB}, x_2^{NB}, \dots, x_N^{NB}\}$ is considered. When aligning the narrowband features with the wideband xHMM, the

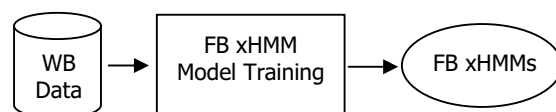


Figure 1: Expansion model training

10.21437/Interspeech.2007-366

original likelihood in Viterbi algorithm has to be modified to account for the missing components like [6]:

$$P_s(\mathbf{x}^{NB}) = \sum_{k=1}^K c_{s,k} \prod_{i=1}^N N(x_i^{NB} | \mu_{s,k,i}, \sigma_{s,k,i}) \quad (1)$$

where s is an xHMM state, K is the number of Gaussian mixture components, c is the mixture component weight, i is the element index of feature vector, and $N(\cdot | \boldsymbol{\mu}, \boldsymbol{\sigma})$ is the Gaussian mixture with mean vector $\boldsymbol{\mu}$ and variance vector $\boldsymbol{\sigma}$. Note that diagonal covariance matrix is assumed.

Step 3: Missing component calculation

Given the alignment of a narrowband feature vector \mathbf{x}^{NB} with the xHMM state s obtained in the previous step, the missing or expanded feature vector components for those vectors are calculated like:

$$x_{N+i}^{EB} = \sum_{k=1}^K p_s(k | \mathbf{x}^{NB}) \mu_{s,k,N+i} \quad \text{for } i = 1 \dots M \quad (2)$$

where

$$p_s(k | \mathbf{x}^{NB}) = \frac{c_{s,k} \prod_{i=1}^N N(x_i^{NB} | \mu_{s,k,i}, \sigma_{s,k,i})}{\sum_{k=1}^K c_{s,k} \prod_{i=1}^N N(x_i^{NB} | \mu_{s,k,i}, \sigma_{s,k,i})} \quad (3)$$

is probability of the mixture component k for the given narrowband feature vector \mathbf{x}^{NB} . The missing components are appended to the narrowband feature vector.

Step 4: Cepstrum calculation

The final step is to calculate MFCC features from the expanded log FB feature vector $\{x_1^{NB}, x_2^{NB}, \dots, x_N^{NB}, x_{N+1}^{EB}, \dots, x_{N+M}^{EB}\}$ by applying the discrete cosine transform (DCT). Furthermore, the usual ASR post-processing can be performed on MFCC features, such as mean removal, and delta and acceleration feature calculation.

3. Experimental setup

3.1. Databases and HMM models

In the monolingual tests, the Aurora 4 WSJ clean-speech training-testing scenario (US English) is used. In this scenario, there is about 15 hours of 16 kHz training data (7000+ utterances) and about 40 minutes of 16 kHz testing data (330 utterances). The database contains read speech and its vocabulary size is 5k.

For multilingual tests, four languages from Speecon database were selected: Spanish, Italian, German, and French.

Speecon was collected in different environments (office, home, car, public place, etc.) using four different microphones. The close talk microphone signals are used in this work. Read sentences and phonetically rich words were used for acoustic model training totaling in above 20 hours of training material for each language. For tests, a selected set of read sentences of about 40 minutes was used for each language. Vocabulary ranged from 1.5k to 2k.

Monophone HMM models with three states are used for both the NB2WB expansion model and the wideband ASR acoustic models. In the multilingual case, a nearly hundred phonemes covering the all five involved languages (four from Speecon plus WSJ) were mapped into a common inventory of 58 phonemes for building the xHMM model. As for the ASR HMM model, the results are reported for two sizes, 64 (small) and 128 (large) Gaussian mixture components.

3.2. Features for ASR HMM and xHMM

Acoustic features for ASR HMM are calculated according to the ETSI advanced front-end standard [7]. This standard provides noise robust MFCC features and it is used to calculate features from the wideband training and testing portions of signals. However, in order to perform a straightforward NB2WB expansion of narrowband signal, log FB features should be used for the xHMM model. These features were obtained using the same ETSI standard right before the DCT calculation step. For wideband signal, the standard calculates 26 log FB features together with a combination of the c_0 feature with the log energy feature (c_0_logE), while for narrowband signal, 23 log FB + c_0_logE features are calculated. Thus, three missing FB features from the high frequencies have to be obtained by NB2WB expansion (i.e. $M=3$ in (2)).

It is well known that MFCC features are more suited for HMM modeling with diagonal covariance matrices than log FB features. To build a better performing xHMM model and still use FB-like features, the mean from each log FB vector was removed and appended as an additional feature to the mean-removed log FB vector (denoted as log FBM); in this way the recovery of the original log FB features is possible when needed. To build the wideband xHMM models, 26 log FBM + Mean + c_0_logE were used as static features augmented by their delta and acceleration versions. Table 1 shows the performance increase obtained on the WSJ task when using log FBM features instead of log FB features. The performance of MFCC features is also shown.

Table 1: Word accuracies for different features

WSJ 16 kHz	MFCC	log FB	log FBM
Word Acc	77.34	50.69	75.76

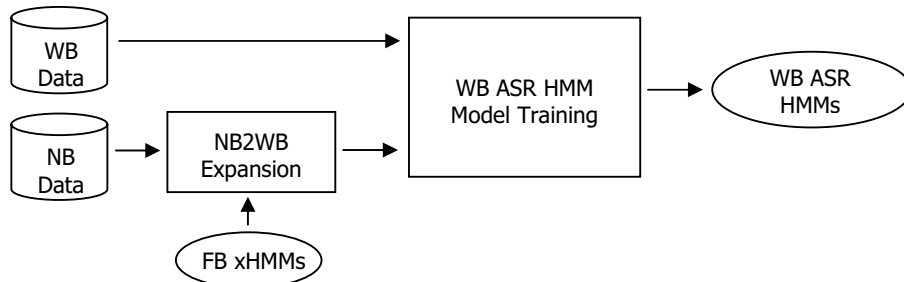


Figure 2: Wideband ASR HMM training using NB2WB expanded data

4. ASR results

4.1. Monolingual tests – US English

In monolingual tests, the Aurora 4 WSJ scenario is used. First, different splits of training data between the wideband and narrowband categories are presented and then the NB2WB expansion algorithm is tested.

4.1.1. Training data splits and baseline results

The training data was split in two parts: WB Data – wideband data, and NB Data – narrowband data (down-sampled from 16 kHz to 8 kHz). Two different splits are considered and are shown in Table 2.

Table 2: Two different splits of WSJ training data

	WB Data	NB Data
Split 1	6.6h (44%)	8.4h
Split 2	2.2h (15%)	12.8h

For each split, the xHMM model is trained using the available wideband data (e.g. in the case of Split 2, 2.2 hours of training material is used). Then, the narrowband data is expanded and used together with wideband data to train the wideband acoustic models for ASR. This process is depicted in Figure 2.

As shown later, the two splits described in Table 2 allow simulating two different HMM model training situations:

- **TR1:** Enough wideband speech data available to train given ASR HMM. The ASR performance will not increase significantly when adding more wideband training data, and thus the objective is that the addition of NB2WB expanded training data causes a minimum (possibly none) decrease of the performance obtained when using only the wideband training data.
- **TR2:** Not enough wideband speech data available to train given HMM. In this case, the addition of NB2WB expanded training data should improve the ASR performance obtained when using only the wideband training data.

Table 3 shows the wideband ASR performance when using different amounts of wideband training data simulating the two training situations. Two different HMM configurations for the ASR acoustic models are shown. The performance of narrowband ASR system is also shown for reference in the last column.

Table 3: Baseline performances for WSJ task

Test →		16k			8k
Train →	8k	---	---	---	100%
	16k	100%	44%	15%	---
HMM →	small (64 m)	76.28	76.18	71.08	71.14
	large (128 m)	80.30	77.34	66.71	75.16

It can be observed that when decreasing the amount of training data to 44% (6.6 hours), the performance for the small ASR HMM configuration does not change significantly when compared to using all training data. In other words, adding more data to the 44% training data does not improve the performance, and thus, neither can be expected an improvement when adding NB2WB expanded training data. This would correspond to the situation TR1 mentioned

previously. Situation is different when considering the large ASR HMMs; adding data to the 44% training data improves the performance from 77.34% to 80.30%, which indicates the TR2 situation. Note that the performance obtained when using all wideband training data provides an upper performance limit for the narrowband to wideband expansion technology.

When using only 15% of the original wideband data (2.2 hours), the performances for both small and large ASR HMM configurations decrease significantly when compared to using the all training data. This indicates a lack of training data (the situation TR2) and thus provides room for improvement when adding the NB2WB expanded data to training.

4.1.2. Tests with NB2WB expanded training data

In the following tests, two approaches for the narrowband to wideband expansion are compared. The first approach consists in performing a simple up-sampling from 8 kHz to 16 kHz, which in the presented case means there is no information added to the three high frequency FB bands. The second approach is the proposed FB xHMM-based expansion approach.

Table 4: Performance of wideband ASR models assuming 6.6 hours (44%) of wideband training data

Test →		WSJ 16 kHz		
Train →	16k	44%	44%	44%
	8k-upsamp-16k	---	56%	---
	8k-xHMM-16k	---	---	56%
HMM →	small (64m)	76.18	72.53	75.93
	large (128m)	77.34	77.15	79.15

Table 4 shows the results for the case when using 44% of wideband training data and 56% are the expanded training data. We can observe that when adding 56% of 8 kHz to 16 kHz up-sampled training data, for small ASR HMM the performance decreases from 76.18% to 72.53%, and for large ASR HMM, the performance practically does not change.

When adding 56% of training data expanded by xHMM, the performance for the small HMM configuration decreases slightly from 76.18% to 75.93%; notice that an improvement is not expected in this case due to the TR1 training situation. For large HMM configuration, the performance increases from 77.43% to 79.15%. This performance is relatively close to 80.30%, the result obtained when using all wideband data for training, and it is better than using 100% of narrowband data for training (75.16% in Table 3).

Table 5: Performance of wideband ASR models assuming 2.2 hours (15%) of wideband training data

Test →		WSJ 16 kHz		
Train →	16k	15%	15%	15%
	8k-upsamp-16k	---	85%	---
	8k-xHMM-16k	---	---	85%
HMM →	small (64m)	71.08	66.01	74.19
	large (128m)	66.71	70.44	76.39

A similar behavior can be observed when using only 15% of wideband training data. Table 5 shows that in this case adding up-sampled data decreases the performance for the small ASR HMM configuration, but it increases the quite low performance of the large ASR HMM configuration from 66.71% to 70.44%. The addition of xHMM expanded training

data increases the performance for both small and large ASR HMM configurations from 71.08% to 74.19% and from 66.71% to 76.39%, respectively.

4.2. Multilingual tests

For multilingual NB2WB experiments, a mix of languages shown in Table 6 was used to train a unique, 58 monophone expansion HMM. This xHMM was then used to expand narrowband features for all involved languages. Notice that no Italian data is used to train xHMM, as it is assumed that no wideband data is available for this language. For Italian wideband ASR HMM training, only the expanded data are used.

Table 6: *Mix of training data for the unique xHMM*

	Spa	Fre	Ger	USEn
Amount	5.4h	6.8h	8.1h	2.2h

Table 7 shows recognition performances for all Speecon tests. For each language, different amounts of wideband data were tested. For example, for German, when using all training wideband material consisting of 24.8 hours of data, the word accuracy of 79.12% is obtained. When using only 8.1 hours of wideband training data, the performance decreases to 71.39%; this would correspond to the TR2 training situation indicating a lack of training data. When adding expanded data, the performance increases to 75.26%. Similar pattern can be observed for all tested Speecon languages; addition of xHMM expanded features into the ASR acoustic model training improves the speech recognition performance. In the case of Italian language it was assumed that there is no wideband data available at all. Still, as it can be observed from the last column of Table 7, the performance of models that use only the expanded training data, 73.19%, is relatively close to the performance of ASR models that use all wideband training data, 75.97%.

5. Conclusions

In this paper, an algorithm for narrowband to wideband expansion of features for robust ASR was presented and tested within single language and multiple language scenarios. The feature expansion was performed by using the Viterbi alignment of narrowband features with a wideband expansion HMM model. The expansion model was trained in a log filter-bank-like domain that allowed a straightforward feature expansion and provided a better ASR recognition performance – and thus a better alignment – than the straight log filter-bank domain.

Two different situations were considered for wideband ASR model training: a) sufficient amount of wideband

training data and b) lack of wideband training data. Experiments showed that the proposed expansion algorithm behaves adequately in these two situations; particularly, it does not decrease the speech recognition performance in the case a) and it improves significantly the recognition performance in the case b).

In multilingual tests, a unique expansion model was trained for four European languages using the Speecon and WSJ databases. In these tests, availability of different amounts of wideband training data was assumed, including no wideband data. In all cases, the wideband ASR models trained by a combination of limited amount of wideband features and narrowband to wideband expanded features performed better than using the limited amount of wideband features alone.

6. Acknowledgements

I would like to thank to my colleagues Yuan-Jun Wei and Yan-Ming Cheng for their help and discussions on the topic.

7. References

- [1] Avendano C., Hermansky H., Wan E. A., “Beyond Nyquist: Towards Recovery of Broad-bandwidth Speech from Narrow-bandwidth Speech”, Proc. Eurospeech 1995, Madrid Spain, pp. 165-168, 1995.
- [2] Cheng Y.M., O’Shaughnessy D., Mermelstein P., “Statistical Recovery of Wideband Speech from Narrowband Speech”, IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, pp. 544-548, October 1994.
- [3] Liao Y.-F., Lin J.-S., Tsai W.-H., “Bandwidth Mismatch Compensation for Robust Speech Recognition”, Proc. Eurospeech 2003, Geneva Switzerland, pp. 3093-3096, September 2003.
- [4] Seltzer M. L., Acero A., “An EM Algorithm for Training Wideband Acoustic Models from Mixed-Bandwidth Training Data”, Proc. IEEE ASRU 2005, pp. 197-202, November 2005.
- [5] Kontio J., Laaksonen L., Alku P., “Neural Network-based Artificial Bandwidth Expansion of Speech”, IEEE Trans. on Audio Speech and Language Processing, vol. 15, no. 3, pp. 873-881, March 2007.
- [6] Barker J., Cooke M., Josifovski L., Green P., “Soft Decisions in Missing Data Techniques for Robust Automatic Speech Recognition,” *ICSLP 2000*, Beijing, 2000.
- [7] ETSI ES 202050, “Speech Processing; Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithm”, ETSI Standard, July 2002.

Table 7: *ASR performances for Speecon languages without and with NB2WB expanded training data*

Test →		German 16 kHz			Spanish 16 kHz			French 16 kHz			Italian 16 kHz		
Train	16k	21.6h	5.4h	5.4h	24.8h	8.1h	8.1h	20.7h	6.8h	6.8h	20.2h	0h	0h
	8k-xHMM-16k	---	---	16.2h	---	---	16.7h	---	---	13.9h	---	---	20.2h
Word Acc →		79.12	71.39	75.26	86.00	77.44	83.73	73.17	64.39	68.99	75.97	N/A	73.19