SPEECH BANDWIDTH EXTENSION USING ARTICULATORY FEATURES

BY

DONGEEK SHIN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Bachelor of Science in Electrical Engineering
in the College of Engineering of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Adviser:

Professor Mark Hasegawa-Johnson

ABSTRACT

In this paper, we present a technique for bandwidth extension (BWE) of a narrow-band (0 - 4 kHz) signal using articulatory features. The proposed technique recovers high-band components (4 - 8 kHz) through Gaussian mixture regression (GMR) on both the acoustic and articulatory features from the X-ray Microbeam (XRMB) speech production database. The Gaussian mixture model (GMM) that is based on acoustic and articulatory features is initialized using k-means and iteratively trained using expectation-maximization (EM) algorithm. BWE experiments were run using data files from different speakers in XRMB as train and test data. Time-frequency plots of speech recovered by different training methods are presented in order to show that articulatory trajectories are helpful in characterizing high frequencied consonants in speech. Finally, we confirm our hypothesis that using GMM with articulation gives better recovery rate is true by performing Student's t-test on SNR data between original and recovered speech.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

CHAPTER 2: GAUSSIAN MIXTURE REGRESSION (GMR)

2.1 Gaussian Mixture Model (GMM)

2.2 Parameter Re-estimation

2.3 Minimum Mean Square Error (MMSE) Criterion

CHAPTER 3: ARTICULATORY TRAJECTORIES

CHAPTER 4: EXPERIMENTS

4.1 Algorithms

4.2 Data

4.3 Training

4.4 Results

CHAPTER 5: DISCUSSION

CHAPTER 6: CONCLUSIONS

CHAPTER 7: REFERENCES

CHAPTER 1

INTRODUCTION

Bandwidth extension (BWE) of low-band signals has been an interesting topic in acoustic engineering. In practice, BWE is frequently performed to improve telephone speech. The bandwidth of a typical telephone speech is limited to around 4 kHz because of the limit in the channel capacity, and thus results in a bad quality for the listeners. In order to avoid the cost of replacing communication hardwares, engineers in speech technology work on methods in BWE to improve speech quality at the receiving end.

There is no standard solution for the problem of BWE of damaged speech. It is because speech signals are complex from their high variability of spectrum with many formant envelopes changing in time. Also, the fact that the spectral data of speech is always different from one speaker to another does not allow hard decision making classifiers to do good jobs as regression models. Previous methods perform BWE using hidden Markov models (HMM) [9], non-negative matrix factorization (NMF) features [7], etc. Also, one technique of BWE was performed using linear prediction (LP) to recover the spectral residues and line spectral frequencies (LSF) [8] to recover the envelopes. Such recovery technique is represented by a linear transformation, $Z_{high} = A \cdot Z_{low}$, where A is the linear transformation matrix from low-band spectrum Z_{low} to high-band spectrum Z_{high} acquired from training LSF models. Similarly, we want to model a function between Z_{high} and Z_{low} but using probability and statistics.

Statistical learning methods are helpful in building their models because it does not assume a deterministic signal as its input. Speech signals are highly variant in time with unique features which can only be effectively analyzed in frequency domain. By learning a Gaussian mixture model (GMM), which is a linear combination of Gaussian probability distributions in acoustic feature space, we are making soft decisions and allowing variability of input speech to be recovered. In other words, we want to

obtain a function $g(\cdot)$ between the low-band and high-band that is not constrained to be linear.

Meanwhile, there has been developments in mapping from articulatory movements to vocal tract spectrum [5] and from spectral data to articulatory features [3], [6] that show the effectiveness of statistical learning methods for mapping problems. Based on the fact that a bijection exists between the acoustic and articulatory data, articulatory movements can be used to build statistical models that performs BWE through a non-linear regression technique. Our aim is eventually obtaining a mapping function $g(\cdot)$ so that $Z_{high} = g(Y)$, where Y is combines articulatory and low-band acoustic features. By demonstrating the effectiveness of including articulatory information in band mapping, applications such as recovering audio bandwidth from dynamic MRI [10] seems possible in the future.

CHAPTER 2

GAUSSIAN MIXTURE REGRESSION (GMR)

From a frequentist perspective, bandwidth extension (BWE) can be performed by collecting an unimaginable number of data sets, and hope that the testing data matches a significant amount of the training data. This would be both time taking and costly in terms of implementation. On the other hand, a Bayesian perspective recovers data based on belief and a relatively small training set needed for BWE using probability distributions to model the dataset. We will construct a probability model of Gaussian mixtures and use it as our regression function.

2.1 Gaussian Mixture Model (GMM)

A Gaussian distribution is a probability distribution that is constructed using two parameters: mean and variance. A multivariate normal distribution is a Gaussian distribution for feature space with dimension greater than 1. It has the probability density function $p_i(x \mid \theta_i)$ where i is an index number as follows.

$$p_i(x \mid \theta_i) = \frac{1}{(2\pi)^{n/2}|\sum|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu)}$$

Here, n is the number of dimensions, μ is the mean, and $\sum$ is the covariance matrix between features.

Gaussian mixture model (GMM) is a collection of individual Gaussians with means and variances. In GMM, weights of Gaussians are introduced as an extra variable. The pdf $p(x \mid \theta)$ of GMM with M Gaussians is simply a collection of those multivariate Gaussian pdfs:

$$p(x \mid \theta) = \sum_{i=1}^{M} \alpha_i p_i(x \mid \theta_i), \quad \sum_{i=1}^{M} \alpha_i = 1$$

where α_i is the mixture weight. The sum of weights go to 1 since probability under the Gaussian function cannot exceed 1.

2.2 Parameter Re-estimation

In order to construct a mixture model based on initial training data, one can start by initializing its parameters using an algorithm rather than picking random points in feature space as means of Gaussians.

K-means algorithm uses Euclidean metric to calculate the cluster points of data. We see that the means of individual Gaussian distributions is under the expression $(x - \mu)$. Since the distribution means and data points are related in difference, it is sensible to initialize our mixture model with k-means clustering that emphasizes distances between the data points and means. Also, k-means is cheap in terms of algorithmic complexity. When c_i, $i \in \{1, 2, ..., N\}$ are the clusters, k-means reestimates the initial cluster points iteratively by minimizing W [2]:

$$W = \sum_{i=1}^{k} \sum_{j=1}^{N} \parallel x_j^{(i)} - c_i \parallel^2$$

where k is the number of clusters and N is the number of features. Whenever W is not minimized, the the centroids of the new clusters are updated as centers and the process of re-estimating clusters is iterated.

The GMM parameters, mean and covariance, are obtained from individual Gaussian estimations for every cluster. The cluster centers are interpreted as the mean vector of our mixture model, and the covariance matrix and prior vector are calculated knowing the number of data points and their distribution in each cluster.

Our initialization of GMM using k-means is a preparation for the expectation-maximization (EM) algorithm. EM algorithm iteratively updates the mixture parameters in search for maximum likelihood [1]. The likelihood function is simply

$$L(\theta \mid X) = \prod_{i=1}^{N} p(x_i \mid \theta)$$

where N is the size of the data set. The likelihood function depends on mixture density parameters θ and describes how well the data fit into our probability model when assuming Gaussian conditions. Thus, maximizing the likelihood function is a way to obtain mixture density parameters that accurately describe our data set.

K-means determines good starting cluster points for EM, and this is useful since EM algorithm reaches its local optimum point of maximum likelihood based entirely on its starting cluster points. Because the likelihood in the parameter space is typically not convex, EM will converge but not necessarily to a global maximum.

According to [1], the parameters for maximum likelihood mixture densities are obtained by maximizing the expectation value of the likelihood function. This algorithm, known as expectation-maximization (EM), calculates the membership of each point in cluster and re-estimates the Gaussian mixture distribution.

The membership weight, or the posterior component probability, is calculated for every data:

$$p(l \mid x_i, \Theta) = \frac{\alpha_i p(x_i \mid \Theta)}{\sum_{i=1}^{N} \alpha_i p(l \mid x_i, \Theta)}$$

If we solve the optimization problem that maximizes our expectation, which is the sum of posterior component probabilities, then, we end up with the following expressions for GMM parameteres, where they are scaled by the posterior probabilities [1].

$$\alpha_l^{new} = \frac{1}{N} \sum_{i=1}^{N} p(l \mid x_i, \Theta)$$

$$\mu_l^{new} = \frac{\sum_{i=1}^{N} x_i p(l \mid x_i, \Theta)}{\sum_{i=1}^{N} p(l \mid x_i, \Theta)}$$

$$\Sigma_l^{new} = \frac{\sum_{i=1}^{N} p(l \mid x_i, \Theta)(x_i - \mu_l^{new})(x_i - \mu_l^{new})^T}{\sum_{i=1}^{N} p(l \mid x_i, \Theta)}$$

where Θ represents the Gaussian parameters of mean, covariance, and weights.

2.3 Minimum Mean Square Error (MMSE) Criterion

A criterion for estimation based on a probabilistic model is required for signal recovery. An estimator $\hat{g}(\cdot)$ that maps low-band to high-band is assumed. The error variance between the original frequency components and recovered frequency components using is minimized to find the minimum mean square error mapping function $\hat{g}(\cdot)$ [3].

$$Z = \hat{g}_{MMSE}(X) = \underset{\hat{g}(\cdot)}{\operatorname{argmin}}(E[(Z - \hat{g}(X))^T(Z - \hat{g}(X))])$$

where Z is the data generated using regression and X is the original data. When we solve the optimization problem, we discover that the recovered data is simply the conditional expectation of that data given our original data.

$$z = E(Z|X = x)$$

CHAPTER 3

ARTICULATORY TRAJECTORIES

Articulatory trajectories spatial coordinate recordings of articulators, which are speech organs in our mouth and vocal tract that have movements corresponding to our speech acoustics. The articulatory feature space thus includes the movement data of lips, tongue, jaw, velum, etc. Because articulatory trajectories are direct physical recordings, they are easily interpreted compared to spectral data, which assumes complex source modeling. The trajectories are recorded using electromagnetic articulography (EMA) that uses sensors to detect the movements while a person is speaking.

Many worked on the acoustic-to-articulatory and articulatory-to-acoustic mapping problems by modelling non-linear functions $X = f(Z)$ and $Z = f^{-1}(X)$ where Z is acoustic data and X is articulatory data. Bandwidth extension using articulatory features essentially deals with a subset of the mapping problem. The mapping of high-frequency components Z_{high} can be based on both articulatory and low-band acoustic information so that $Z_{high} = g(Y)$ where $Y = [Z_{low};\ X]$. [5] discusses how applying maximum likelihood estimation (MLE) on GMM-based mapping from articulatory to vocal tract spectrum shows best mapping performance giving minimum spectral distortion.

CHAPTER 4

EXPERIMENTS

4.1 Algorithms

Bandwidth extension (BWE) is performed by a GMM-based mapping using articulatory features. The GMM training sequence is shown in Fig 4.1. The controllable variables that we allow in this experiment are number of mixture components, bandwidth of test data, re-estimation type, and type of features.

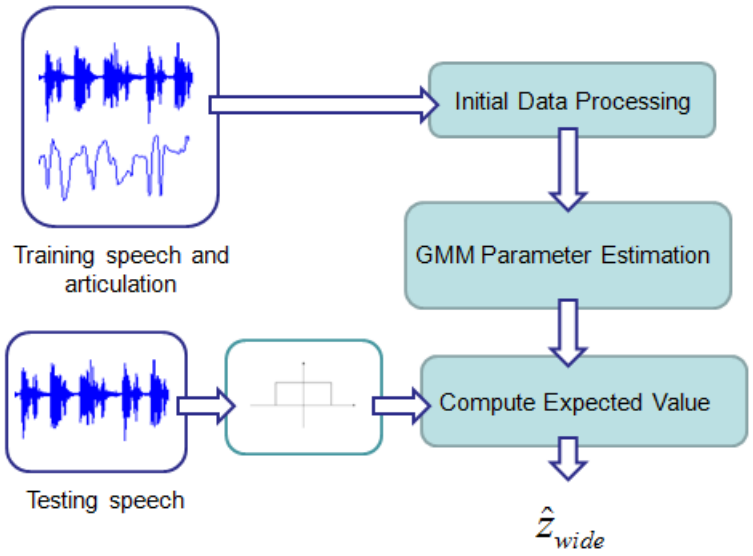

Fig. 4.1. High-level diagram of BWE using GMR

Initial data processing includes obtaining spectral data from time-domain acoustics, and resampling and resizing of articulatory and acoustic features. Every spectrum data was calculated using a hamming window with distance of 6866 μs, which is the window length of articulatory data.

K-means initialization of GMM parameters generates the mean μ_i simply as the vector of newly estimated cluster points. If $W_i = \{w_1, w_2, ...\}$ is a non-empty set of features that exist as data points in cluster $i \in \{1, 2, ..., k\}$, we generate the priors as follows.

$$\alpha_i = \frac{|W_i|}{\sum_{j=1}^{k} |W_j|}$$

Also, if $W_i = [w_1, w_2, ...]$, the covariance matrix calculated for each cluster.

$$\Sigma_i = \frac{W_i W_i^T}{N} - \mu_i \mu_i^T$$

where μ_i is the mean for cluster i.

Gaussian distributions may collapse into a delta function where the covariance matrix near singular. We thus add a floor variance to ensure numerical stability.

$$\Sigma_i^{new} = \Sigma_i + \varepsilon I_{n \times n}$$

Our expectation-maximization algorithm is applied for our contatenated acoustic and articulatory feature matrix $Y = [Z_{low};\ X]$ for every cluster.

$$\alpha_l^{new} = \frac{1}{N} \sum_{i=1}^{N} p(l \mid y_i, \Theta)$$

$$\mu_l^{new} = \frac{\sum_{i=1}^{N} x_i p(l \mid y_i, \Theta)}{\sum_{i=1}^{N} p(l \mid y_i, \Theta)}$$

$$\Sigma_l^{new} = \frac{\sum_{i=1}^{N} p(l \mid y_i, \Theta)(yi - \mu_l^{new})(y_i - \mu_l^{new})^T}{\sum_{i=1}^{N} p(l \mid y_i, \Theta)}$$

We use MMSE for our recovery criterion. Converting our GMM distribution and its parameters into conditional distribution, mean, and covariance, we obtain the following expression for minimum mean square error estimation for high-band If we trained our GMM based on only on acoustics:

$$\hat{z}_{high,MMSE} = \sum_{i=1}^{k} p(l \mid z) \left[\mu^i_{Z_{high}} + \Sigma^i_{Z_{high} Z_{low}} (\Sigma^i_{Z_{high} Z_{low}})(z - \mu^i_{Z_{low}}) \right]$$

where $\Sigma^i_{Z_{high} Z_{low}}$ is the covariance matrix between low-band and high-band components for cluster i of our trained GMM.

If our GMM is with articulation, our expected means are conditioned upon Y.

$$\hat{z}_{high,MMSE} = E(Z_{high} | Y = [z;\ x])$$

$$= \sum_{i=1}^{k} p(l \mid y) \left[\mu^i_{Z_{high}} + \Sigma^i_{Z_{high} Y} (\Sigma^i_{Z_{high} Y})(y - \mu^i_Y) \right]$$

4.2 Dataset

X-ray microbeam (XRMB) database, developed at the University of Wisconsin, includes simultaneous recordings acoustic and articulatory data. The speech files in XRMB atabase are from a male or female speaker reading a complete paragraph, a sentence, series of digits, and six to eight words separated by silences. In XRMB, articulatory data were obtained by microbeam detection for pellets (T1, T2, etc.) glued to different places in articulators [4].

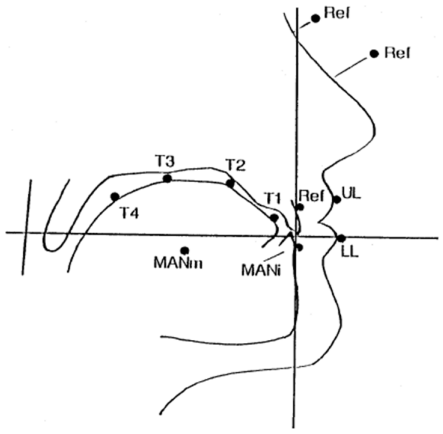

Fig. 4.2. Locations of pellets in mouth

TABLE 4.1. TYPES OF PELLETS FOR XRMB TECHNIQUE

Pellet Type	Pellet Name	Number	Nominal Sampling Rate
reference	MAX(i)	3	40
mandibular	MAN(i)	2	40
upper lip	UL	1	40
lower lip	LL	1	160
ventral tongue	T1	1	80
mid-tongue	T2, T3	2	80
dorsal tongue	T4	1	80

The XRMB database includes *.xyd* files, which are sets of continuous spatial coordinates of all pellets in time. For *.xyd* files, all pellet data are resampled to 6.866 ms (approximately 146 Hz) for the purpose of length normalization. Each *.xyd* file comes in pair with a sound wave file, which is the speech acoustics, and has coordinate difference values between the reference pellets and the 8 others. Each pellet can record its movement in two spatial dimensions (x and y). Hence, the articulatory data we use in our experiments have a total of 16 independent trajectory data.

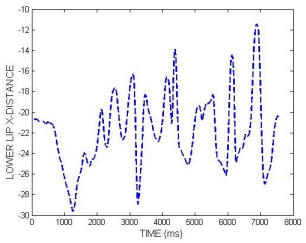

Fig. 4.3. X-coordinates of pellet on lower lip

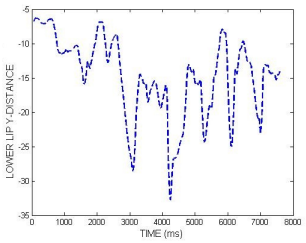

Fig. 4.4. Y-coordinates of pellet on lower lip

The articulatory trajectories, as described in Fig 4.3. and Fig 4.4., are recordings of movements of articulators and are continuous. All coordinates are encoded in units of microns (10^{-3} mm).

4.3 Training

Speech from speaker JW15 of X-ray Microbeam (XRMB) speech production database is used to train our model. The training set has total of 79 words from sound files that have continuous sentences, fast digit reading, and citation word files. We did not discriminate and included all types of training data. The words in training data does not overlap the words in testing data.

XRMB database recorded the number 1000000 for articulator coordinates that are considered bad because of artifacts in the experiment. We excluded all frames of 1000000 when concatenating the training data.

The dimensions of our feature space is set to 16 for the articulatory trajectories and 75 for the spectral data. When only training with the acoustic file, the dimension of the Gaussian mixture is thus 75.

4.4 Results

The figures shown are the spectrograms for the data file TP021, in which a male speaker lists the words “country understand silk sense hall” separated by silences.

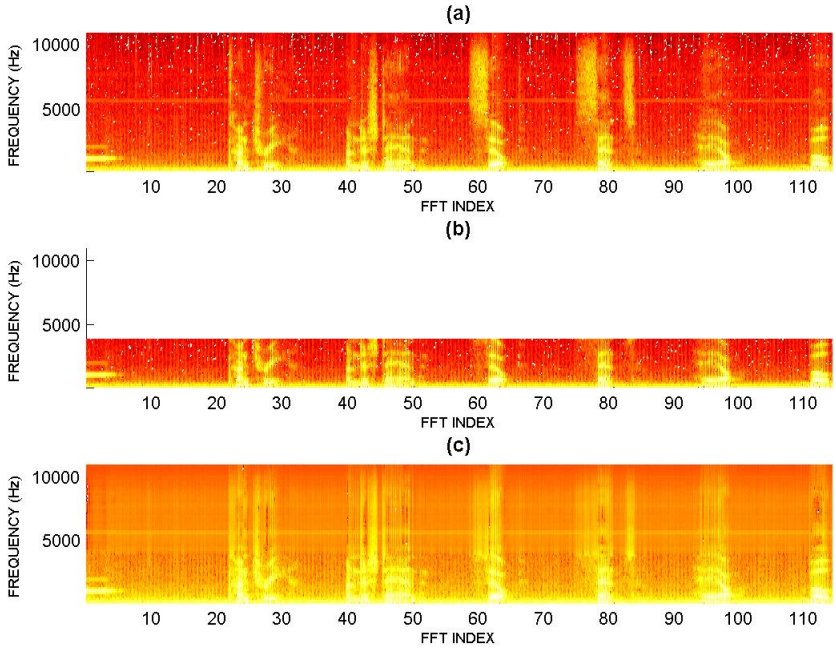

Fig. 4.5. (a) original speech, (b) bandlimited speech, cut number = 45, (c) recovered speech using 16 mixtures without articulatory features

Fig 3.1. shows the recovery procedure using GMM without articulation. The last spectrogram of Fig 4.5. describes how the high frequency components of the 5 words are recovered. Since there are 75 frequency indeces,

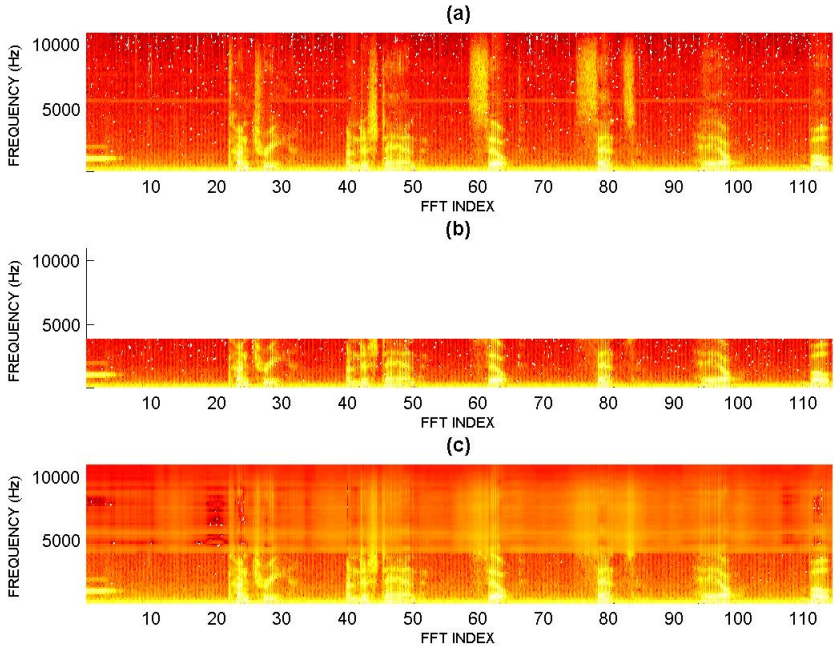

Fig. 4.6. (a) original speech, (b) bandlimited speech, cut number = 45, (c) recovered speech using 16 mixtures with articulatory features

One improvement in recovery that is seen from the spectrograms from Fig. 4.5. to Fig 4.6. is the estimation of the temporal positions of high frequency components. For example, for the word "country", the original GMM with acoustic feature training has recovered frequency envelopes nearly constant in time.

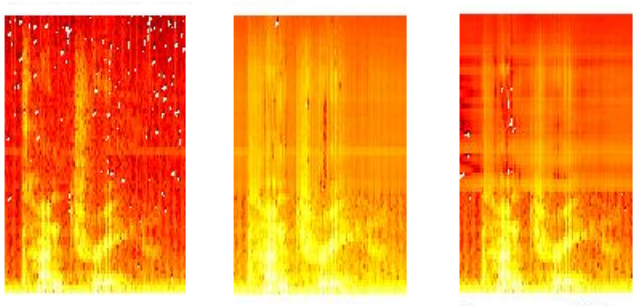

Fig. 4.7. Spectrogram for "country". Left to right: original, recovered without articulatory, recovered with articulatory

Fig 3.3 displays the spectrogram of a specific word "country" recovered using different features. Our GMM without articulation recovers the first frequency burst at c with series of harmonics that are not present in the original signal. However, the burst well estimated with decaying harmonics by GMM with articulation.

We also tested extreme cases when the original band only exists up to 2 kHz. Words like school which start with a strong s thus a evident high frequency component. Thus, when the frequency components of speech are severely cut, there is not even a sign of recovery of fricatives when we use GMM with only acoustic data training. This plot does not study recovery, since the frequencies are blurred, but shows how articulatory data is able to detect frequencies even with minimal data.

The signal-to-noise (SNR) ratio, where noise is the spectral distortion, is computed from our spectrogram data.

$$SNR = 10\log_{10}\left(\frac{\sum_{i=1}^{N}(Z_{recovered})^2}{\sum_{i=1}^{N}(Z_{real} - Z_{recovered})^2}\right)$$

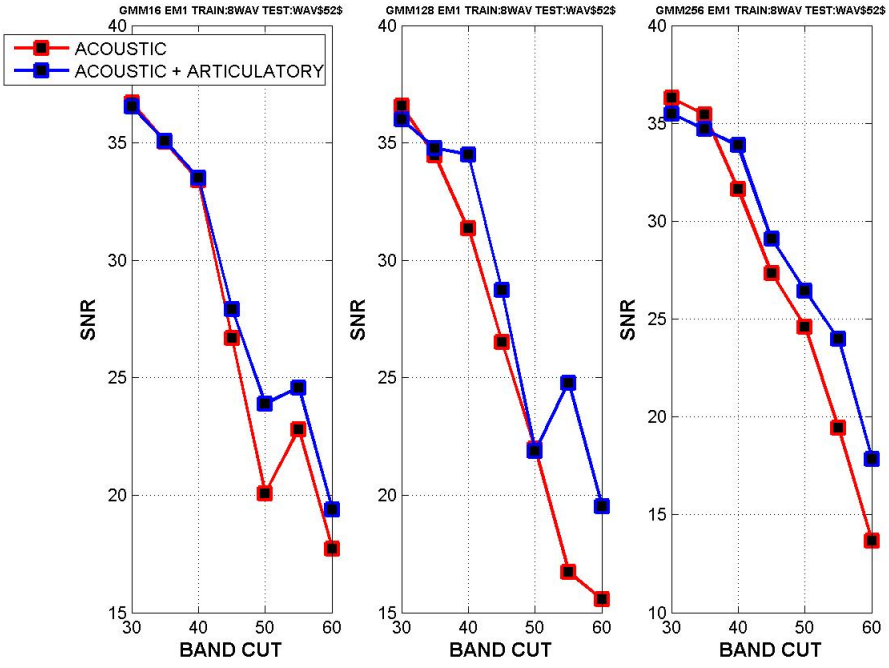

Fig. 4.8. SNR in spectral values for different bandlimit frequencies

The plots in Fig. 3.3. show how signal-to-noise ratio (SNR) changes with different band cuts for three mixture types: 16, 128, and 256 Gaussians. Comparing SNR for original GMM and GMM with articulation for any case, GMM with articulation is always higher in SNR when recovering a narrow-band signal, which has frequency components up to 4 kHz $\cong$ 40 band cut number. For band cut number higher than 45, GMM that used both acoustic and articulatory data has higher SNR with a larger difference.

We claim that GMM trained with articulation recovers low-band (0 - 4 kHz) data with less spectral distortion. Using the Student's t-test, we are able to determine whether our hypothesis is true or not. We test whether our calculated t, our test statistic, follows or approximates a Student's t-distribution given null hypothesis. The SNR data are

$SNR_{acou} = [24.6070\ 19.4176\ 13.6594\ 16.7371\ 20.0670\ 22.7933\ 17.7184]$

$SNR_{arti} = [26.4491\ 23.9760\ 17.8348\ 24.7735\ 23.8926\ 24.5767\ 19.3880]$

and are from the SNR data for GMMs trained on the first 8 audio file and tested with speaker JW15's 52th audio file. The calculation of t is supported by other variables.

$$S_{xx} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{n_1 + n_2}{n_1 - n_2}}$$

$$s_1 = \sqrt{Var(SNR_{acou})},\ s_2 = \sqrt{Var(SNR_{arti})}$$

$$df = n_1 + n_2 - 2,\ t = \frac{(\mu_{articulatory} - \mu_{acoustic})}{S_{xx}}$$

According to our calculations, degree of freedom when df equals 12, t is 2.0297. We validate our initial hypothesis that GMM with articulation recovers speech better than GMM without articulation. According to the t-table, for 90 percent confidence interval, $t = 1.782$. Thus, all GMM mappings from articulatory information recovers high-band with less spectral distortion than mappings using only acoustic data.

CHAPTER 5

DISCUSSION

Fricatives, such as letters s and z, and plosives, such as p, occupy largely the high frequency spectrum above 5 kHz since they are unvoiced and described mostly by noise. Joint training of acoustic and articulatory features is effective especially in recovering them. When GMM is only trained with acoustics, because even in the training data fricatives are high-frequencied, the Gaussian mixture cannot distinguish between silences and fricatives when recovering a severly damaged test file. If the letters s or z dominate in our speech, then there would have been better recovery rate with only the acoustic model, but this is typically not the case in normal speech.

Also, Fig 3.2. shows although GMM with articulatory information approximates high frequency components very well, we noticed that it generates some smearing effects in the recovered speech. However, the relative amplitude difference between the energy of speech and the smearing is very high as seen from our SNR plots.

We cannot neglect error inherent in the dataset. As the XRMB handbook describes, raster hops, frames where the microbeam detects and records the wrong pellet, are evident in several places in data. Even though this data collection error does not dominate the entire set, such inconsistency in the original data ripples throughout the entire experiment. Acquiring a better recorded articulatory data set would improve the statistical model since they are closer to 'ground truth'.

Our current model can improve by selecting a better feature representation than spectral vectors for articulatory data. A feature set that describes speech with excitation signals but still is a compressed version of spectral data would be useful. However, it is important that the features are invertible, since we are extending the bandwidth of audio not only for analytical purposes but also practical purposes of improving sound quality.

CHAPTER 6

CONCLUSIONS

A bandwidth extension technique for narrow-band (0 - 4 kHz) speech using articulatory features was proposed. We chose our mapping function from low-band to high-band as $Z_{high} = g(Y)$ and Y includes both low-band components and articulatory trajectories. We trained our Gaussian mixture model, using k-means and expectation-maximization algorithm, and performed Gaussian mixture regression based on minimum mean square error criterion on the low frequency components to recover the high frequency using articulatory trajectories as our feature.

Our SNR results of recovered speech show that the presence of articulatory data recovers the high frequency components with less spectral distortion. Our t-test on SNR data of testing data and original data shows that using GMM with articulation gives better recovery rate. Also, the spectrograms tell us that articulatory trajectories are powerful in characterizing high frequencied consonants such as plosives and fricatives. For cases where the bandwidth of the damaged signal is extremely low, articulatory features maybe first employed to simply detect the temporal positions of high frequency components.

CHAPTER 7

REFERENCES

[1] J. Bilmes. "A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models," ICSI, 1998.

[2] P. Smaragdis. "Machine Learning for Signal Processing," CS 598 Course Notes, August 2011.

[3] Y. Özbek, M. Hasegawa-Johnson, and M. Demirekler. "Estimation of Articulatory Trajectories Based on Gaussian Mixture Model (GMM) With Audio-Visual Information Fusion and Dynamic Kalman Smoothing," IEEE Trans. Audio, Speech, Language Processing, 2011.

[4] J. R. Westbury. "X-ray Microbeam Speech Product Database User's Handbook," University of Wisconsin, 1994.

[5] T. Toda, A. W. Black, and K. Tokuda. "Mapping from Articulatory Movements to Vocal Tract Spectrum with Gaussian Mixture Model for Articulatory Speech Synthesis", Workshop on Speech Synthesis, 2004.

[6] T. Toda, A. W. Black, and K. Tokuda. "Acoustic-to-Articulatory Inversion Mapping with Gaussian Mixture Model", 8th International Conference on Spoken Language Processing, 2004.

[7] D. Bansal, B. Raj, P. Smaragdis. "Bandwidth Expansion of Narrowband Speech Using Non-Negative Matrix Factorization," Interspeech, 2005.

[8] S. Chennoukh, A. Gerrats, G. Mie, and R. Sluijter. "Speech enhancement via frequency bandwidth extension using line spectral frequencies," ICASSP, 2001.

[9] G. Chen and V. Parsa, "HMM-based frequency bandwidth extension for speech enhancement using line spectral frequencies," ICASSP, 2004.

[10] S. Narayanan, K. Nayak, S. Lee, A. Sethy, and D. Byed. “An approach to real-time magnetic resonance imaging for speech production”, The Journal of the ASA, 2004.