STOCHASTIC MODELING OF SPECTRAL ADJUSTMENT FOR HIGH QUALITY PITCH MODIFICATION[1]

Alexander Kain (kain@cse.ogi.edu)

Center for Spoken Language Understanding (CSLU)
Oregon Graduate Institute of Science and Technology
P.O. Box 91000, Portland, OR 97291-1000, USA
http://cslu.cse.ogi.edu

Yannis Stylianou (yannis@research.att.com)

AT&T Labs - Research
Speech and Image Processing Lab
180 Park Ave, Florham Park, NJ 07932-0971, USA
http://www.research.att.com

ABSTRACT

We present a new algorithm for adjusting the magnitude spectrum when the fundamental frequency (F_0) of a speech signal is altered. The algorithm exploits the correlation between F_0 and the magnitude spectrum of speech as represented by line spectral frequencies (LSFs). This correlation is class-dependent, and thus a broad classification of the input is achieved by a Gaussian mixture model (GMM). The within-class dependencies of LSFs on F_0 values are captured by constructing their joint probability densities using a series of GMMs, one for each speech class. The proposed system is used for post-processing the pitch modified signal. Perceptual tests showed that the addition of this post-processing system improves the naturalness of the pitch modified signal for large pitch modification factors.

1. INTRODUCTION

Concatenative text-to-speech synthesizers generate speech by piecing together small units from a recorded speech database and then performing a series of signal processing methods to smooth concatenation boundaries and to match the desired prosodic targets (such as speaking speed and pitch contour) accurately. One of the important modifications is adjusting the fundamental frequency (F_0). However, large modification factors lead to a perceptible decrease in speech quality. One of the reasons for this degradation is the assumption of a constant magnitude spectrum. However, the opposite has been shown to be true: In [1], an increase of F_0 was observed to cause a vowel boundary shift or a vowel height change. Additionally, an analytical study showed that the overlap of natural vowels in formant scatter plots can be normalized using F_0 information. In [4], it was found that speakers generally increase the first formant as they increase F_0.

These results suggest that the speech spectrum must be altered in a specific way during pitch modification. Such an approach was taken in [5], where the spectrum was modified by a stretched difference vector of a codebook mapping. However, a shortcoming of this method is that only three ranges of F_0 (high, middle, and low) are encoded, assuming a very simple evolution of the spectrum with changing F_0.

In this work, we propose an algorithm that can predict the average evolution of the spectrum over all naturally occurring F_0 values by means of a stochastic model trained on a speaker database.

The paper is organized as follows: In Section 2, we describe the statistical analysis and modeling techniques, followed by an overview of the spectral adjustment algorithm in Section 3. Section 4 and 5 discuss the implementation and evaluation of the system.

2. STATISTICAL ANALYSIS AND MODELING

2.1 Analysis

To learn about the correlation between F_0 and spectral envelope, we synchronously computed F_0 values and bark-scale warped line spectral frequencies (LSFs) over the center 50ms (if not limited by a phoneme boundary) of every voiced phoneme of a large, single speaker database. The measurements were stored as feature vectors in a database. The decision to employ LSFs is motivated by their good interpolation and coding properties, as demonstrated in [3]. Additionally, the bark-scale warping effects

[1] This work was carried out during a summer internship at AT&T Labs in Florham Park, NJ.

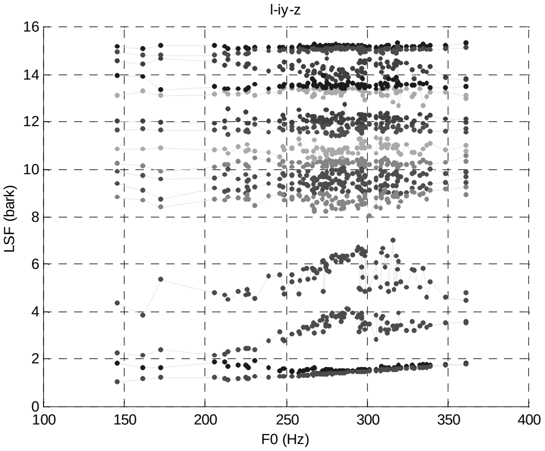

Figure 1. LSFs versus F_0 for all occurrences of /iy/ in the context of /l/iy/z/ as in "please" for a female speaker. 16 bark corresponds to approximately 8kHz.

a frequency weighting that is in agreement with human perception.

Statistical analysis on the feature database revealed that there are significant correlations between F_0 and LSFs as shown by analysis of variance tests. However, these correlations are dissimilar for different speech classes. An example of the spectral evolution with F_0 for a particular class is shown in Figure 1. In this example, it can be observed that the frequencies of the first line spectral pair are increasing, while their distance is decreasing with higher F_0 values. If this pair corresponds to the first formant (F_1), then it can be said that F_1's frequency is increasing and its bandwidth is narrowing with an increase of F_0.

2.2 Modeling the joint density by a GMM

For every speech class, we model the statistical dependency of F_0 and LSFs using a Gaussian Mixture Model (GMM) whose parameters have been estimated on the joint density of F_0 and LSFs [2].

A GMM models the probability distribution of a statistical variable z as the sum of Q multivariate Gaussian functions,

$$p(z)=\sum_{i=1}^{Q}\alpha_i N(z;\mu_i,\Sigma_i),\ \sum_{i=1}^{Q}\alpha_i=1,\ \alpha_i\geq 0 \qquad (1)$$

where $N(z;\mu,\Sigma)$ denotes a normal distribution with mean vector μ and covariance matrix Σ, and α_i denotes the prior probability of class i. The parameters of the model (α,μ,Σ) can be estimated using the Expectation Maximization (EM) algorithm.

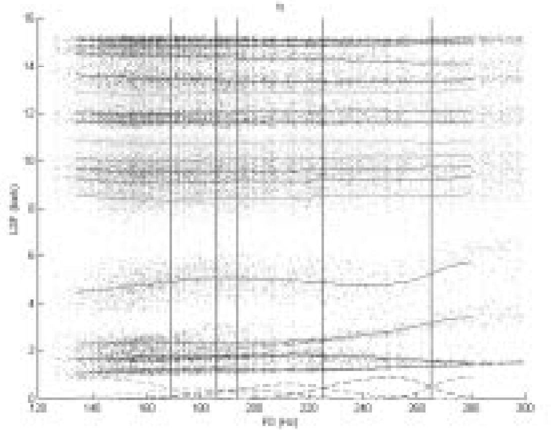

Figure 2. Mapping GMM trajectories (lines) for the class /iy/ (no context) of a female speaker, superimposed on actual data (points) from the feature database. The dashed lines on the bottom represent the normalized posterior probability for six different classes within the mapping GMM, whereas the vertical lines represent boundaries of regions of maximum likelihood of a particular mixture component.

For a particular speech class, let $x=[x_1\ x_2\ \cdots\ x_N]$ represent a sequence of estimated F_0 values and $y=[y_1\ y_2\ \cdots\ y_N]$ the simultaneously estimated LSF feature vectors. The goal is to compute a mapping function F that minimizes the mean squared error

$$\varepsilon_{mse}=E\left[\|y-F(x)\|^2\right] \qquad (2)$$

where E denotes expectation.

To model the joint density, we vertically join x and y to form

$$z=\begin{bmatrix}x\\y\end{bmatrix} \qquad (3)$$

and estimate GMM parameters (α,μ,Σ) for the density $p(z)$, which is the joint density $p(x,y)$.

A locally linear mapping function that attempts to minimize the mean squared error between predicted and target vectors is the regression

$$\begin{aligned}F(x)&=E[y\mid x]=\int y\ p(y\mid x)\,dy\\&=\sum_{i=1}^{Q}h_i(x)\left[\mu_i^y+\Sigma_i^{yx}(\Sigma_i^{xx})^{-1}(x-\mu_i^x)\right]\end{aligned} \qquad (4)$$

where

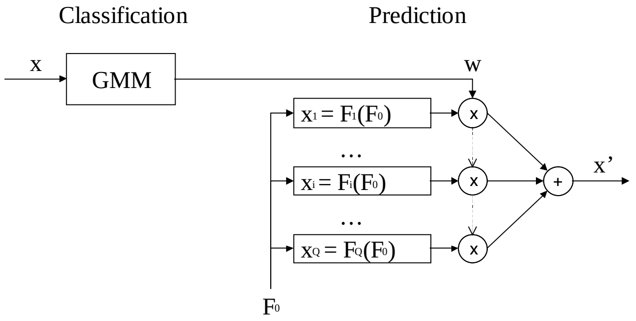

Figure 3. Spectral adjustment algorithm overview. The input spectrum *x* is classified and a corresponding weighting vector *w* is transmitted. A weighted sum of nonlinear functions F predict the target spectrum *x'* from F_0. The prediction functions are realized by mapping GMMs.

$$h_i(x) = \frac{\alpha_i N\left(x;\mu_i^x,\Sigma_i^{xx}\right)}{\sum_{j=1}^{Q}\alpha_j N\left(x;\mu_j^x,\Sigma_j^{xx}\right)} \tag{5}$$

with

$$\Sigma_i = \begin{bmatrix} \Sigma_i^{xx} & \Sigma_i^{xy} \\ \Sigma_i^{yx} & \Sigma_i^{yy} \end{bmatrix}, \quad \mu_i = \begin{bmatrix} \mu_i^x \\ \mu_i^y \end{bmatrix} \tag{6}$$

is the posterior probability that the i^{th} Gaussian component generated *x*, calculated by application of Bayes' theorem.

An example of predicted LSFs over the range of F_0 for the phoneme /iy/ can be seen in Figure 2. It is noteworthy that, by estimating the joint density, decision boundaries are not evenly spread over the F_0 range, but instead are adjusted to capture relevant differences in the LSFs as well.

3. SYSTEM DESCRIPTION

In this section, we describe how the proposed algorithm computes the spectral adjustment. First, the necessary initial classification of the current speech sound is carried out by a "classifying" GMM. Then, a weighted sum of nonlinear prediction functions, realized by a series of "mapping" GMMs, yields the final target magnitude spectrum.

3.1 Classification

Let *x* represent an input LSF vector. In this stage, the algorithm "softly" classifies the input vector by determining the posterior probabilities of a GMM with *Q* mixture components (see Figure 3).

Training of the classifying GMM is possible in a supervised or an unsupervised mode. In supervised training, the parameters for the GMM classifier are directly computed from the distribution of LSFs per phonetic class. In this case, the number of mixture components equals the number of phonetic classes. In unsupervised training, an expectation maximization (EM) algorithm is employed to model the input space probabilistically, with any desired number of mixture components.

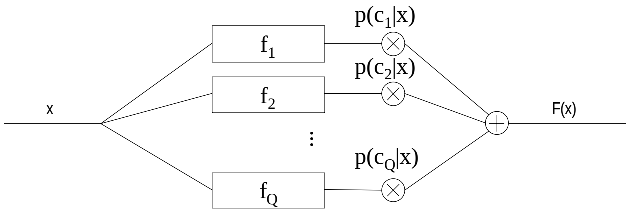

Figure 4. A mapping GMM implements a nonlinear function as a sum of class-dependent linear transfomations weighted by the posterior probability that the input was generated by the distribution of that class.

The posterior probabilities are normalized to form a weighting vector which is passed to the next stage.

3.2 Prediction

A sum of nonlinear mapping functions, weighted by the weighting vector of the classifier, predicts the resulting vector *x'* from the current F_0 value. There is one mapping function for every class of the classifying GMM (see Figure 3).

Each mapping function is realized by a mapping GMM. Such a GMM implements a nonlinear function as a number of class-dependent linear transformations weighted by the posterior probability that the input was generated by the distribution of that class (see Figure 4). The parameters of the linear functions and posterior probability estimator were determined as discussed in the previous section.

4. SYSTEM IMPLEMENTATION

F_0 values and LSF vectors were extracted from labeled voiced speech of a female speaker in the manner described in Section 2.1. The speech sampling rate was 16kHz and the order of the LSFs was set to 16. The final feature database contained 308,027 vectors. This is equivalent to approximately 4 hours of speech, since each vector was determined from a segment up to 50ms long.

We implemented the classifying GMM with 27 voiced phonetic classes in supervised mode. As an indicator of the performance of the classifying GMM we measured the average phoneme recognition accuracy at 62% (see Figure 5). Most

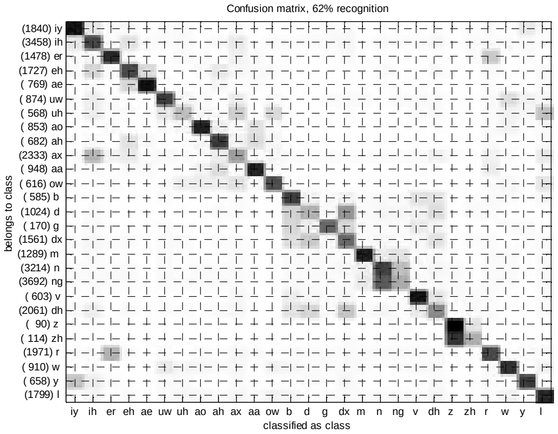

Figure 5. Confusion matrix of the GMM classifier in supervised mode. The numbers on the left represent total occurences in the database. A darker box represents a higher value.

classification errors were made in phonetically similar classes (such as /n/ and /ng/, /z/ and /zh/). Mixture models with 32 and 64 classes were also constructed in unsupervised mode by the EM algorithm. In this mode, a class-labeling of the feature database was achieved through a partitioning according to the maximum likelihood of the classifying GMM.

For each (phonetic or arbitrary) class, a mapping GMM was constructed as outlined in Section 2.2. It was found that six mixture components modeled the spectral adjustment satisfactorily.

When estimating parameters for both the mapping and the classifying GMMs (in unsupervised mode), the means of the models were initialized to the values of the codebook of a binary splitting vector quantizer of the feature database. Then the EM algorithm was run until the increase in likelihood was below a set threshold.

For voiced sounds, speech was modified by an overlap-add method which employed filters whose frequency responses were designed to change the estimated spectral envelope to the target spectral envelope. This was achieved by setting the filter numerator coefficients to the estimated all-pole model coefficients and the filter denominator to the target all-pole model coefficients. Unvoiced sounds were not modified.

5. SYSTEM EVALUATION

We carried out a preference test to assess the performance of the system. Five sentences underwent modifications by a PSOLA method. In addition, the proposed pitch modification algorithm was run. The F_0 modification factors were: 0.5, 0.8, 1.2, 1.5, and 2.0. Subject listened to 25 pairs of two sentences, one of which was produced by PSOLA, the other by PSOLA and the proposed post-processing algorithm. Listeners were then asked to indicate which utterance they preferred in terms of speech quality.

In a small study with 5 listeners, an average preference of 66% for the proposed method was found. Scores for the 0.5 modification factor in isolation were 84%, and for the 2.0 modification factor 76%. For the modification factors 0.8, 1.2, and 1.5, listeners commented that they had trouble hearing a difference between the pairs. This is consistent with PSOLA's adequate performance for moderate modification factors.

6. CONCLUSION

We presented a new algorithm for adjusting the magnitude spectrum as a post-processing step to the pitch modification of a speech signal. The algorithm exploits the correlation between F_0 and the magnitude spectrum of speech by predicting the spectral adjustment using a stochastic model trained on a large speaker database. This correlation depends on the speech class, and thus a GMM was employed for a broad classification. For each speech class, a mapping GMM was constructed to model the evolution of LSFs over the range of F_0 values. Perceptual tests demonstrated that listeners prefered the post-processing of pitch modified speech for large modification factors.

ACKNOWLEDGMENTS

The authors thank the volunteers of the perceptual tests for their assistance.

REFERENCES

[1] T. Hirahara, "On the Role of Fundamental Frequency in Vowel Perception," The Second Joint Meeting of ASA and ASJ, November 1988.

[2] N. Kambhatla, *Local Models and Gaussian Mixture Models for Statistical Data Processing*, Ph.D. thesis, Oregon Graduate Institute, Portland, OR, January 1996.

[3] K. K. Paliwal, "Interpolation Properties of Linear Prediction Parametric Representations," Proceedings of EUROSPEECH, pp. 1029-32, September 1995.

[4] A. K. Syrdal and S. A. Steele, "Vowel F1 as a Function of Speaker Fundamental Frequency", 110th Meeting of JASA, vol. 78, Fall 1985.

[5] K. Tanaka and M. Abe, "A new fundamental frequency modification algorithm with transformation of spectrum envelope according to F0," ICASSP vol. 2, pp.951-954, 1997.