SPECTRAL NORMALIZATION EMPLOYING HIDDEN MARKOV MODELING OF LINE SPECTRUM PAIR FREQUENCIES *

Bryan L. Pellom and *John H.L. Hansen*
Robust Speech Processing Laboratory
Duke University, Box 90291, Durham, NC 27708-0291
http://www.ee.duke.edu/Research/Speech bp@ee.duke.edu jhlh@ee.duke.edu

ABSTRACT

This paper proposes a spectral normalization approach in which the acoustical qualities of an input speech waveform are mapped onto that of a desired neutral voice. Such a method can be effective in reducing the impact of speaker variability such as accent, stress, and emotion for speech recognition. In the proposed method, the transformation is performed by modeling the temporal characteristics of the Line Spectrum Pair (LSP) frequencies of the neutral voice using hidden Markov models. The overall approach is integrated into a pitch synchronous overlap and add (PSOLA) analysis/synthesis framework. The algorithm is objectively evaluated using a distance measure based on the log-likelihood of observing the input (or normalized input) speech given Gaussian mixture speaker models for both the input and desired neutral voice. Results using the Gaussian mixture model formulated criteria demonstrate consistent normalization using a 10 speaker database.

1. INTRODUCTION

The ability to transform or modify the speech characteristics of speakers with differing voice qualities towards a desired neutral or "normalized" voice has several fundamental applications. It is well known, for example, that speaker variability such as accent [1], stress and emotion [2], and physical variabilities such as vocal tract length can seriously degrade the performance of modern speech recognizers. Hence, by developing algorithms which compensate or normalize for such variations, the sensitivity of speech recognizers to such perturbations can be reduced. In addition, other studies have attempted to impart voice personality traits as part of overall speech synthesis systems [3, 4, 5]. Finally, improved normalization and speech analysis/synthesis techniques will ultimately lead to new methods for generating more natural sounding synthetic speech of various vocal qualities.

Several previous studies have considered normalizing voices using vector quantization (VQ). In these approaches, VQ-codebooks consisted of pitch, power and spectral parameters for both the input and desired neutral voice. A mapping procedure such as Dynamic Time Warping is used to associate the parameter spaces between the two speaking styles. Normalization is then carried out by performing linear prediction (LP) analysis on the input speech and applying a codebook mapping to transform the spectral parameters. A more sophisticated approach was also considered by integrating the spectral normalization into a pitch synchronous overlap and add (PSOLA) framework. Although the mapping between the source and desired normalized voice was also learned by a DTW process, the underlying acoustical classes were determined using unsupervised clustering of the feature parameters.

In this paper, we consider a new analysis/synthesis approach to the problem of speaker voice normalization. In particular, hidden Markov models are employed to model the time evolution of the Line Spectrum Pair (LSP) frequencies of a desired neutral voice. The usage of LSP parameters is motivated by several reasons. First, LSP parameters correlate well to formant location and bandwidth structure. Second, LSP parameters have been shown to exhibit good interpolation properties. This has made LSP parameters well suited for speech coding applications [6]. Finally, previous studies such as [7] have illustrated that LSP parameters encode a fair degree of speaker specific information. In this paper we formulate a normalization procedure which independently models the desired neutral voice characteristics for applications such as improved speaker-independent recognition. Therefore, a set of identical training utterances spoken by both the source and desired neutral voice is not required in order to learn the acoustical mapping (in contrast to related previous studies where a DTW based learning phase was required).

2. ALGORITHM FORMULATION

2.1 Modeling Procedure

The spectral characteristics of the desired neutral voice are modeled via Line Spectrum Pair frequencies. Here, it is assumed that the training speech is phonetically labeled and the phoneme sequence has been time-aligned to each training utterance. The segmentation procedure can either be accomplished by hand-labeling or by an automated procedure such as described in [8]. During the modeling phase, the speech data from the training corpus is analyzed on a frame-by-frame basis using a 5 msec frame rate and a 25

*This material is based upon work supported in part by a National Science Foundation Graduate Research Fellowship.

msec analysis window length. For each short-time segment, a set of 10 Line Spectrum Pair frequencies are computed from a 10^{th} order LP analysis as outlined in [9]. The observation vectors are then divided into training tokens for the set of phoneme labels which occur in the training corpus. From the tokenized parameter set, the model parameters for a 5-state, single-mixture hidden Markov model are iteratively estimated for each phoneme using the Baum-Welch method.

Given the set of monophone HMMs for the desired neutral voice, an LP error residual is associated with each state of the Markov chain. Hence, each state consists of a prototypical excitation sequence (source) in addition to an LSP mean vector (filter). The excitation sequence is associated with each state by performing pitch synchronous analysis of the training speech. For each pitch synchronous analysis waveform, a corresponding LSP parameter vector and 10^{th} order LP error residual is computed. Next, we choose the analysis waveform whose corresponding LSP vector is closest to the state mean LSP vector. Here, the Inverse Harmonic Mean (IHM) distance [10] is computed. This distance metric weighs mismatch in formant location more heavily than mismatch in formant bandwidth structure. Thus, for an analysis waveform whose LSP vector is $\vec{x}$, the distance between $\vec{x}$ and an HMM state mean LSP vector, $\vec{y}$, is calculated by,

$$d_{IHM} = \sum_{i=1}^{P} w_i (x_i - y_i)^2 \qquad (1)$$

where w_i is a weighting factor defined as,

$$w_i = \left[\frac{1}{x_i - x_{i-1}} + \frac{1}{x_{i+1} - x_i} \right] \qquad (2)$$

with $P = 10$, $x_0 = 0$ and $x_{P+1} = \pi$. Given the analysis waveform whose LSP vector is closest to the HMM state mean LSP vector, we associate the LP error residual from the corresponding analysis waveform with the HMM state. This procedure is repeated for each state in the model.

Finally, the median pitch (F_0) for the desired neutral voice is estimated from the training speech. In summary, the desired neutral voice characteristics are modeled by (1) a set of monophone HMMs consisting of state dependent mean LSP parameters and prototypical source characteristics, and (2) the overall median pitch value of the desired voice characteristic.

2.2 Normalization Procedure

Assuming the desired neutral speech model parameters are known, the normalization is outlined as follows. The pitch-mark locations within the input utterance are first determined using the time-domain PSOLA technique [11]. For each pitch-mark, an analysis waveform is obtained by windowing a short-time segment of the input speech centered about the pitch-mark location (the width of the window is chosen as twice the local pitch period). Because voiced speech conveys a greater degree of speaker specific qualities for speech recognition, we consider normalization of voiced speech segments only. Hence, the analysis waveform is unaltered in the case of unvoiced speech while a normalized analysis waveform is desired in the case of voiced speech.

The normalization procedure for voiced phonemes is illustrated in Figure 1. Here, the voiced phoneme segment boundaries (shown in Figure 1 as τ_i and τ_{i+1}) are assumed known. With knowledge of the phoneme boundaries, the state dependent mean vectors of the HMM are uniformly positioned within the boundaries. Note that the mean vector components can be thought of as encoding the prototypical spectral properties of the desired neutral voice. Using cubic spline interpolation, a normalized LSP parameter vector is obtained from the HMM mean vectors for the time-instant of the analysis pitch-mark.

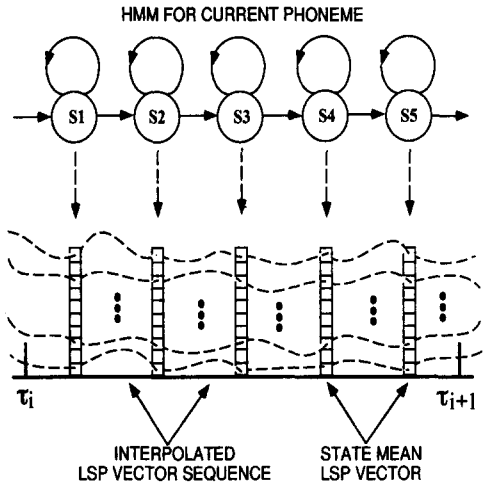

Figure 1: Illustration of LSP sequence determination.

In a similar fashion, the excitation sequence (LP error residual) associated with each HMM state is used to determine a normalized residual. Here, linear interpolation of the state dependent residuals is performed to obtain the normalized residual at the pitch-mark time instant. Finally, a normalized analysis waveform is obtained by first transforming the LSP parameter vector into a corresponding LP coefficient vector. The analysis waveform is then recovered by filtering the normalized error residual with the all-pole filter. Note here that the overall frame energy of the output waveform is adjusted to match that of the input.

With knowledge of the overall median pitch of the input speech (determined using a pitch tracking algorithm), the synthesis pitch-mark locations are positioned such that the overall median pitch of the output normalized voice is equivalent to that of the neutral voice. This is equivalent to scaling the pitch of the output waveform by a constant. In this study, the prosodic structure and time-scale of the input waveform are not modified (although the PSOLA framework could be used to compensate for fast or slow speaking styles). Finally, the normalized speech is recovered using the pitch synchronous overlap and add method.

3. ALGORITHM EVALUATION

3.1. Experimental Database

Speech from 10 adult male speakers of American English was collected using a head-mounted microphone. The speech was sampled at a rate of 8 kHz. Table 1 summarizes the age, height, and median pitch value for each speaker considered in this study.

Speaker Label	Age	Height	Median Pitch (F0)
A	36	1.83 m	103 Hz
B	24	1.70 m	139 Hz
C	27	1.73 m	110 Hz
D	29	1.91 m	129 Hz
E	24	1.84 m	125 Hz
F	26	1.80 m	133 Hz
G	29	1.85 m	130 Hz
H	24	1.75 m	127 Hz
I	25	1.78 m	137 Hz
J	39	1.78 m	97 Hz

Table 1: Summary of speakers used in spectral normalization experiments.

The speech data consists of a set of 19 isolated monosyllabic word tokens (repeated 3 times) and 18 continuous speech utterances. The text transcription for each utterance was selected from the TIMIT database in order to ensure a well-rounded phonetic balance. The data was phonetically time-aligned to the spoken phoneme sequence using an HMM based speech segmentation algorithm.

3.2. Training Procedure

Using the 18 utterances and 19 isolated word tokens, a set of 5-state left-to-right monophone HMMs were estimated so that a desired neutral voice could be obtained (i.e., voice A, B, ..., I, J shown in Table 1). Here, the speech training data was analyzed every 5 msec using a 25 msec Hanning window. For each analysis frame, an observation vector consisting of 10 Line Spectrum Pair frequencies was calculated. In particular, 6 iterations of the Baum-Welch re-estimation procedure was used to determine the parameters of each monophone HMM. Using the procedure outlined in Sec. 2.1, an LP error residual was associated with each state in the Markov model.

3.3 Speaker Modeling for Algorithm Evaluation

It has been shown that Gaussian mixture models (GMMs) are useful for modeling speaker identity [13]. As a result, we consider assessing the performance of the proposed algorithm using GMM speaker models. Consequently, we represent each voice S by a Gaussian mixture density ϕ_s. The mixture density is represented as a weighted sum of M multivariate Gaussian densities where each mixture component is defined by a mean vector, $\vec{\mu}$, covariance matrix, Σ, and mixture weight, c. Hence the probability of observing a single speech vector at time t, $\vec{x}(t)$, given a speaker model, ϕ_s, is expressed by,

$$p\left(\vec{x}(t) \mid \phi_s\right) = \sum_{i=1}^{M} c_i b_i\left(\vec{x}\right). \tag{3}$$

where $b_i(\cdot)$ is the i^{th} multivariate Gaussian density.

Let $\vec{X}_s = \{\vec{x}_s(1), \vec{x}_s(2), \cdots, \vec{x}_s(T)\}$ represent a sequence of T observation vectors obtained from an input speech utterance and ϕ_s and ϕ_n represent the GMMs of the original input and general neutral voice respectively. Then the log-likelihood of observing $\vec{X}_s$ given the input and desired neutral voice models can be expressed by,

$$\log \lambda = \log \frac{p\left(\vec{X}_s \mid \phi_n\right)}{p\left(\vec{X}_s \mid \phi_s\right)} = \log \left\{ \frac{\prod_{t=1}^{T} p\left(\vec{x}_s(t) \mid \phi_n\right)}{\prod_{t=1}^{T} p\left(\vec{x}_s(t) \mid \phi_s\right)} \right\}. \tag{4}$$

In other words, the measure indicates how well the input voice scores to the desired neutral voice model relative to how well the input voice scores to it's own model. In this study, a frame-normalized variant of Equation 4 is computed by dividing by the observation count, T, of each utterance.

Noting the above formulation, the continuous speech utterances in the database were analyzed every 10 msec using a 25 msec Hanning window. The parameter set and the number of mixture densities were chosen based on results obtained in [13]. In particular, silence sections were removed from each utterance and observation vectors consisting of 20 mel-cepstral coefficients were calculated. A total of 32 Gaussian densities were used to model each voice and 10 iterations of the EM algorithm were used to estimate the voice model parameters.

3.4 Continuous Speech Evaluation

In this study, we are interested in normalizing the spectral characteristics of a set of input speech utterances (produced by speakers of differing vocal qualities) towards a desired neutral voice. We first evaluate the proposed algorithm by considering a single speaker from Table 1 as possessing the desired neutral voice characteristic. To examine the effectiveness of the proposed algorithm, the speech utterances from the remaining speakers are systematically normalized. Using the log-likelihood ratio measure, the parametric distance of the input speech (before and after normalization) can be compared to that of the desired neutral voice.

Noting the aforementioned procedure, 18 continuous utterances from 9 speakers ($9 \cdot 18 = 162$ utterances) were normalized for the desired neutral voice. The log-likelihood scores obtained from each utterance were averaged in order to generate an overall measure of closeness to the desired spectral characteristics. The results of this evaluation are shown in Figure 2. Here, we see the average log-likelihood score before and after normalization for each possible desired neutral voice scenario. It is apparent that the proposed algorithm increases the log-likelihood of the input speech towards that of the desired neutral voice. In fact, the strong

negative values of the log-likelihood before normalization indicate the small probability that the input speech was generated by a neutral-like voice. After normalization, however, we see that the positive values of the log-likelihood measure confirm that the probability that the processed speech was generated by a neutral-like voice has improved. Overall, we see that the average log-likelihood increases from -3.55 prior to normalization to +1.63 after normalization.

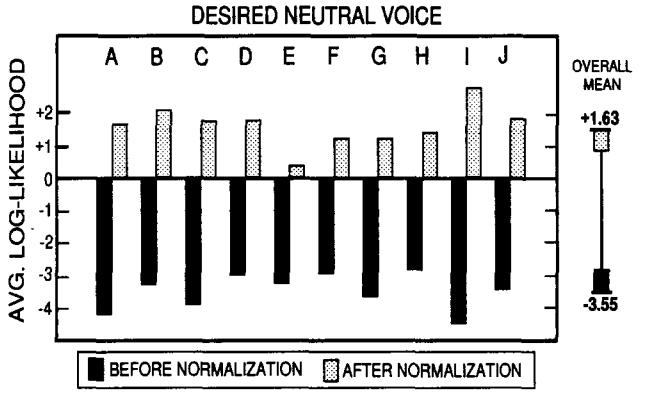

Figure 2: Improvements in average log-likelihood score of speech utterances to desired neutral voice after processing by proposed spectral normalization algorithm.

In light of the results obtained in Figure 2, a second evaluation was performed in order to assess whether the proposed normalization successfully produces speech with spectral characteristics that match closest to only that of the desired neutral voice. In other words, the normalization procedure may generate spectral characteristics which are closer to the desired neutral voice, but at the same time even closer to an undesired voice. To evaluate this scenario, we use the same normalization setup from the first evaluation, but modify the scoring procedure. In particular, we compare the log-likelihood of the normalized speech against GMMs for all voices listed in Table 1. Normalized utterances are considered correctly classified as the desired neutral voice if the log-likelihood of the desired neutral voice is greater than the log-likelihood of the remaining voices in the speaker set. This procedure is analogous to speaker identification [13].

Results of the second evaluation are shown in Table 2. Here we see, for example, that 96% of the utterances normalized to match the spectral characteristics of voice H were correctly classified as being produced by H and not some other voice. In general, we see that the proposed algorithm successfully directs the spectral characteristics in 9 of the 10 cases with the poorest performance resulting from normalizations towards voice E. It is suggested that such a method would prove to be useful for limiting the impact of speaker variability due to accent, emotion, or stress for speech and speaker recognition applications.

4. CONCLUSIONS

In this paper a hidden Markov model based spectral normalization approach was formulated. In the proposed method,

DESIRED NEUTRAL VOICE (SEE TABLE 1)

A	B	C	D	E
100%	100%	100%	100%	42%
F	G	H	I	J
97%	97%	96%	100%	100%

Table 2: Percent of speech utterances correctly classified as the desired neutral voice after processing by proposed spectral normalization algorithm.

the spectral characteristics of a desired neutral voice are modeled using HMMs derived from Line Spectrum Pair frequencies. The overall approach was integrated into a pitch synchronous analysis/synthesis framework which allows joint normalization of spectral and excitation characteristics. A Gaussian mixture model based criteria was formulated to assess the performance of the proposed spectral normalization algorithm. Using this criteria, the average log-likelihood of the input speech parameters was shown to increase from -3.55 (before normalization) to +1.63 (after normalization) indicating consistent normalization of spectral characteristics towards that of a desired neutral voice. Such a scheme could be an effective method for neutralizing input speech for improved speaker-independent recognition.

References

[1] L. M. Arslan, J. H. L. Hansen, "Language Accent Classification in American English," *Speech Communication*, vol. 18, pp. 353-367, August 1996.

[2] J. H. L. Hansen, S. E. Bou-Ghazale, "Robust Speech Recognition Training via Duration and Spectral-Based Stress Token Generation," *IEEE Trans. on Speech and Audio Processing*, vol. 3, pp. 415-421, September, 1995.

[3] S. Bou-Ghazale, J.H.L. Hansen, "Generating Stressed Speech from Neutral Speech using a Modified CELP Vocoder," *Speech Communication*, vol. 20, pp. 93-110, November 1996.

[4] M. Abe, S. Nakamura, K. Shikano, H. Kuwabara, "Voice conversion through vector quantization," *ICASSP-88* , pp. 655-658, 1988.

[5] H. Valbret, E. Moulines, J.P. Tubach, "Voice Transformation using PSOLA technique," *Speech Communication*, vol. 11, pp. 175-187, 1992.

[6] J. Crosmer, T. Barnwell, "A low bit rate segment vocoder based on line spectrum pairs," *ICASSP-85* , vol. 1, pp. 240-243, 1985.

[7] L. Chi-Shi, M. Lin, W. Wang, H, Wang, "Study of Line Spectrum Pair Frequencies for Speaker Recognition," *ICASSP-90* , pp. 277-280, 1990.

[8] F. Brugnara, D. Falavigna, M. Omologo, "Automatic Segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication*, vol. 12, pp. 357-370, 1993.

[9] F. Soong, B.-H. Juang, "Line Spectrum Pair and Speech Data Compression," *ICASSP-84* , vol 1., pp. 1.10.1-4, 1984.

[10] R. Laroia, N. Phamdo, N. Farvardin, "Robust Efficient Quantization of Speech LSP Parameters using Structured Vector Quantizers," *ICASSP-91*, pg. 641-644, 1991.

[11] E. Moulines, J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, pp. 175-205, 1995.

[12] T. Kitamura, M. Akagi, "Speaker Individualities in Speech Spectral Envelopes," *J. Acoust. Soc. of Japan*, vol. 16, no. 5, pp. 283-288, 1995.

[13] D. Reynolds, R. Rose, "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83.