Combining HMM and Weighted Deviation Linear Transformation for Highband Speech Parameter Estimation

Hwai-Tsu HU*(a) and Chu YU†, Members

SUMMARY A hidden Markov model (HMM)-based parameter estimation scheme is proposed for wideband speech recovery. In each Markov state, the estimation efficiency is improved using a new mapping function derived from the weighted least squares of vector deviations. The experimental results reveal that the performance of the proposed scheme is superior to that combining the HMM and Gaussian mixture model (GMM).

key words: wideband recovery, HMM-based parameter estimation, weighted deviation linear transformation

1. Introduction

Telephone speech sounds thin and muffled because the transmission bandwidth is limited below 3.4 kHz. A straightforward solution for improving speech quality is to adopt a wideband network to augment the channel bandwidth and allocate more bits for transmission. However, such an idea appears impractical due to the amount of legacy networks already installed. Fortunately, the missing bandwidth can still be partially recovered from narrowband speech via signal processing techniques [1]–[4].

The basic idea behind wideband recovery lies in the assumption that the spectra in the lowband and highband regions maintain certain relationships. Hence part of the lost highband components are retrievable by performing lowband-to-highband parameter estimation followed by a resynthesis process. Although further improvement is possible by comprising additional information while conducting speech coding [5], this inevitably increases the transmission bandwidth. In this letter, we only focus on how to exploit the HMM for efficient parameter estimation.

2. Experimental Data

We adopt 16 kHz wideband speech to extract the parameters for analysis. The lowband (defined as 0–3.5 kHz) component is acquired by lowpass filtering the wideband speech with a cutoff frequency of 3.5 kHz, followed by 2-to-1 downsampling. This results in an 8 kHz signal that can be treated as narrowband speech. To simulate the effect of the telephone link effect, we further pass this speech signal through a 128-point bandpass FIR filter resembling the telecommunication sending and receiving characteristics [6]. Also, by means of signal processing techniques, we can obtain another down-sampled 8 kHz signal that possesses the exact spectral content in the highband (defined as 3.5 k–7.5 kHz) range. The spectral envelopes of these two bands are characterized using 10th and 8th order line spectrum pair (LSP) parameters [7], termed \{\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{10}\} and \{\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_{8}\} respectively, which are obtained from linear prediction (LP) analysis of the low- and high-band signals. In particular, the gain factor \(y\) for the highband signal is defined as

\[
y = \log_{10} \left( \frac{\sum_{k=1}^{8} \hat{a}_k \hat{R}(k)}{\sum_{k=1}^{10} \hat{a}_k \hat{R}(k)} \right) \quad (1)
\]

where \(\hat{a}_k\)'s are the LP coefficients derived from the autocorrelation sequence \(\hat{R}(k)\)'s of the lowband signal. Analogously, \(\hat{a}_k\)'s and \(\hat{R}(k)\)'s are from the highband signal.

Twenty minutes of speech samples taken from the TIMIT database [8], which is a corpus of American English of different genders and dialects, are used. We analyze the wideband speech on a frame basis. Each frame covers an interval of 30 ms with 50% overlap, thus resulting in a total of 80000 frames.

3. Highband Spectral Parameter Estimation

3.1 HMM-Based Approach

Let the input and output parameters be expressed in vector form as \(x = [r_1, r_2, \ldots, r_{10}]^T\) and \(y = [\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_8, y]^T\), respectively. Here our attention focuses on the estimation efficiency of the intended approach rather than the appropriate-ness of the selected parameters. Hence no other parameters are included in the input vector, despite that the inclusion of additional acoustic features was known to be beneficial to the parameter estimation [3], [4]. In [3], a HMM was utilized to estimate the highband parameters. The output was formulated as

\[
\hat{y}(m) = \sum_{i=1}^{N_i} P(S_i(m) | X(m)) E(y(m) | S_i(m), X(m)) \quad (2)
\]

where \(X(m)\) denotes the feature vector sequence observed up the \(m\)th frame, i.e. \(X = \{x(1), \ldots, x(m)\}\). \(S_i(m)\) signifies the \(i\)th Markov state at frame \(m\), and \(N_i\) corresponds to the
number of states. Using Bayes’ rule, the conditional probability \( P(S_i(m) \mid X(m)) \) was obtained by

\[
P(S_i(m) \mid X(m)) = \frac{P(S_i(m), X(m))}{\sum_{j=1}^{N_t} P(S_j(m), X(m))}.
\]

(3)

The joint probability \( P(S_i(m), X(m)) \) could be computed in a recursive manner as in [3].

In the original formulation, the input vector \( x(m) \) was assumed to have no effect on \( E(y(m) \mid S_i(m), x(m)) \). Thus, the expectation estimate becomes

\[
E(y(m) \mid S_i(m), x(m)) = E(y(m) \mid S_i(m)) = \mu_{i(o)}^y
\]

(4)

where \( \mu_{i(o)}^y \) denotes the mean output vector of state \( S_i \). This approach can be regarded as a kind of codebook mapping (CM), since the lowband-highband relationship is described by a codebook entry. For easy distinction the output vector derived by substituting Eq. (4) into Eq. (2) is denoted as \( \hat{y}_{HMM-CM}(m) \).

Note that the state probability density function (pdf) is generally approximated using a weighted sum of a set of Gaussian distributions, namely the GMM. We can certainly calculate the conditional expectation using the GMM estimator [2], which is performed as

\[
E(y(m) \mid S_i(m), x(m)) = \sum_{j=1}^{Q} h(i,j)(x(m)) [\mu_{i(o)}^y + \Sigma_{i(o)}^y (\Sigma_{i(o)}^{xx})^{-1} (x(m) - \mu_{i(o)}^{xx})],
\]

(5)

where

\[
h(i,j)(x(m)) = \frac{\lambda_{(i,j)} N(x(m); \mu_{i(o)}^{xx}, \Sigma_{i(o)}^{xx})}{\sum_{k=1}^{Q} \lambda_{(i,k)} N(x(m); \mu_{i(o)}^{xx}, \Sigma_{i(o)}^{xx})}
\]

\( \lambda_{(i,j)} \) denotes the prior probability of the \( j \)th Gaussian component associated with the \( i \)th state. \( Q \) represents the number of mixture components. \( N(x(m); \mu_{i(o)}^{xx}, \Sigma_{i(o)}^{xx}) \) stands for a multivariate Gaussian distribution with mean vector \( \mu_{i(o)}^{xx} \) and covariance matrix \( \Sigma_{i(o)}^{xx} \). On the other hand, \( \Sigma_{i(o)}^{xx} \) is a submatrix associated the covariance matrix \( \Sigma_{i(o)}^{xx} \) deduced from the paired vectors in form of

\[
z = [x^T, y^T]^T.
\]

All the involved GMM parameters can be estimated using the well-known expectation-maximization (EM) algorithm [9]. Applying Eq. (5) to Eq. (2), we have an alternative estimation named \( \hat{y}_{HMM-GMM}(m) \).

3.2 Weighted Deviation Linear Transformation (WDLT)

Equations (4) and (5) presents two possible approaches of transforming the lowband parameters to the highband parameters. Another approach also proven to be quite efficient is the linear mapping [1], which has the following form:

\[
y(m) = \frac{M}{\sum_{i=1}^{M} \alpha_i} \hat{V}_i(x(m)) \quad \text{with} \quad \sum_{i=1}^{M} \alpha_i = 1
\]

(6)

where \( M \) denotes the number of space partitions and \( \alpha_i \) is the weighting factor with respect to the matrix \( \hat{V}_i \) that carries out the mapping. In light of the above concept, we can formulate the parameter estimation in state \( S_i \) as

\[
E(y(m) \mid S_i(m), x(m)) = \mu_{i(o)}^y + \hat{V}_i(x(m) - \mu_{i(o)}^x)
\]

(7)

Notice that the vector pair corresponding to a particular codebook entry, namely \( (\mu_{i(o)}^x, \mu_{i(o)}^y) \), still plays an important role in our mapping function. This feature is inherited from the codebook mapping. Moreover, the multiplicative matrix \( \hat{V}_i \) is applied to the offset vector, \( x(m) - \mu_{i(o)}^x \), instead of the input vector \( x(m) \). Let \( \hat{d}_{i(o),m}^y = x(m) - \mu_{i(o)}^x \) and \( \hat{d}_{i(o),m}^y = y(m) - \mu_{i(o)}^y \). The matrix \( \hat{V}_i \) can be obtained by minimizing the sum of weighted squared errors, i.e.

\[
\sum_{m=1}^{N_f} \omega_m (\hat{d}_{i(o),m}^y - \hat{V}_i \hat{d}_{i(o),m}^x)^2, \quad \text{for} \quad N_f \quad \text{frames gathered in the training set. This leads to the consequence that}
\]

\[
\hat{V}_i = \left( \sum_{m=1}^{N_f} \omega_m \hat{d}_{i(o),m}^x (\hat{d}_{i(o),m}^x)^T \right)^{-1} \sum_{m=1}^{N_f} \omega_m \hat{d}_{i(o),m}^x (\hat{d}_{i(o),m}^y)^T.
\]

(8)

As for the weighting factor, we simply treat the GMM pdf, namely \( \sum_{k=1}^{Q} \lambda_{(i,k)} N(x(m); \mu_{i(o)}^{xx}, \Sigma_{i(o)}^{xx}) \), as \( \omega_i(x(m)) \) since it reflects the significance of the observed vector to the state \( i \). Since all training vectors participate in the derivation, Eq. (8) tends to emphasize the influence of vectors with larger pdf’s but ignores the smaller ones. By incorporating Eqs. (7) and (8) into (2), we finally get

\[
\hat{y}_{GMM-WDLT}(m) = \sum_{i=1}^{N_f} P(S_i \mid X(m))[\mu_{i(o)}^y + \hat{V}_i(x(m) - \mu_{i(o)}^x)].
\]

(9)

4. Performance Evaluation

The performance of the aforementioned parameter estimation techniques is evaluated using a spectral distortion measure defined as

\[
D_S(y, \hat{y}) = \frac{1}{N_f} \sum_{f=1}^{N_f} \int_{0}^{0.5f_s} \frac{1}{0.5f_s \omega_s} \left( 20 \log_{10} \left| \frac{g_f}{A_f e^{j \omega_f}} \right| \right)^2 \, d \omega_s
\]

(10)

where \( \omega_s \) is the sampling frequency. \( A_f(e^{j \omega_f}) \) and \( \hat{A}_f(e^{j \omega_f}) \) denote the original and recovered LP spectra computed from the output vectors \( y \) and \( \hat{y} \), respectively. The variable \( g_f \) and \( \hat{g}_f \) are the associated gain factors derived from the actual \( y \) and estimated \( \hat{y} \), i.e. \( g_f = 10^b \) and \( \hat{g}_f = 10^\hat{b} \).
Figure 1 presents the averaged distortion measures obtained from HMM-CM, HMM-GMM, and HMM-WDLT for the training set with $N_s = 2^4, 2^5, 2^6$. The results indicate that, for the HMM-CM and HMM-GMM, raising either the state number or the mixture component number in each state can lower the spectral distortion. Of course, the gain in the estimation accuracy is at the cost of additional computation and memory requirement. Also, under the HMM framework, both the GMM and WDLT estimators outperform the CM originally proposed in [3]. In our experiments the WDLT even surpasses the GMM. However, unlike the declining trend observed in the GMM cases, the spectral distortions attained by the WDLT remain roughly at the same level regardless how many Gaussian components are employed to calculate the weighting factors as required by Eq. (8). Such a consequence is conceivably due to that the refinement of the weighting factors does not result in a significant change in the multiplicative matrix $V_{(i)}$. Eventually, a single Gaussian pdf is good enough to derive $V_{(i)}$.

5. Conclusion

An HMM-based scheme for highband speech parameter estimation has been proposed. While the HMM is adopted to model the statistical dependencies among speech frames, the lowband-to-highband parameter estimation in each Markov state is attempted using the GMM and WDLT, both of which are proven to be better than the CM. As the WDLT works more efficiently than the GMM, using HMM and WDLT is thus the suggested combination to achieve optimal performance in highband parameter estimation.

Acknowledgments

This work was supported by the National Science Council, ROC, under Grant NSC96–2221–E–197–020–MY2.

References