RELIABLE BANDS GUIDED SIMILARITY MEASURE FOR NOISE-ROBUST SPEECH RECOGNITION

Zhang, Bo and Peng, Gang and Wang, William S-Y

Department of Electronic Engineering, City University of Hong Kong, Hong Kong

ABSTRACT

Under noisy conditions, due to the redundancy of speech signal, there are some spectral bands (Reliable Bands) whose local SNR’s are high enough to be used effectively by a recognizer. A novel, phonetically motivated Reliable Bands Guided similarity measure (RBG measure) is proposed in this study. It has the following features. Firstly, for reference spectrum, frequency bands which have larger absolute energy or sharper spectral peaks are marked as reliable bands. They are to be given more weight than the other bands in the definition of the RBG measure. Secondly, within each reliable band, similarity between formant positions and formant shapes of test spectrum and reference spectrum is explicitly modelled. Lastly, the measure can automatically emphasize spectral bands whose amplitudes change abruptly, which normally contain more reliable dynamic features of the speech signal. Both the RBG measure and the Parallel Model Combination (PMC) method are tested on a speaker-independent, continuous Mandarin digit string recognition task, under 15 noisy conditions. Noises are drawn from the NOISEX92 database. The RBG measure shows an average 4.22% word accuracy score below the PMC method above 0 dB. However, it outperforms the PMC method by 8.82% at 0 dB. More importantly, the RBG measure does not rely on accurate background noise modeling, which is a difficult task in itself.

1. INTRODUCTION

Noise-robust speech recognition is a challenging task. Various techniques have been proposed to combat this problem. These techniques can be roughly divided into four categories (see [5] for discussions on the methods listed below): (1) Some methods are based on robust speech representations, e.g., the RASTA-PLP analysis and the Cepstral Mean Normalization; (2) Some methods employ noise-robust distance measures, e.g., the Cepstral Projection measure and the Factored Spectrum similarity measure; (3) Some methods model the background noise and, based on the noise model, estimate the embedded clean speech, or predict the property of the noisy signal. This category includes the spectral subtraction methods, the state-based Wiener filtering, the Joint Cepstral Compensation, the Stochastic Matching, and the model composition and decomposition techniques; (4) Some methods rely on partial spectro-temporal information for recognition. The noise masking method and the missing data techniques belong to this category. Among all these techniques, the model composition method, Parallel Model Compensation (PMC) [4], shows more effectiveness under various testing conditions.

In this study, a Reliable Bands Guided similarity measure (RBG measure) is proposed. It belongs to the category (4) above. Motivations behind the RBG measure are: (a) Under noisy conditions, due to the redundancy of speech signal, there are some spectral bands whose local SNR’s are still high enough. Within these bands, noisy spectrum is either quite close to the original spectrum, or is just a broadened version of the latter. So more weight should be given to these bands when designing a similarity measure, (b) Formants information plays a more important role in speech perception than the whole spectral shape does. Thus, operating in log-spectral domain, our similarity measure will explicitly model the similarity between formant positions and formant shapes of test spectrum and reference spectrum.

2. GENERATION OF TEMPLATE SET BY THE MONTE-CARLO METHOD

Suppose the emitting probability density function (pdf) of state \( s_j \) of an HMM is modelled by

\[
b_j(x) = \sum_{m=1}^{M} w_{jm} N(x; \mu_j^m, \Sigma_j^m),
\]

where \( x \) is a cepstral observation vector, \( w_{jm} \) is the weight of the \( m^{th} \) mixture within the total \( M \) mixtures, \( N(x; \mu_j^m, \Sigma_j^m) \) is the multivariate Gaussian distribution of the \( m^{th} \) mixture whose mean vector is \( \mu_j^m \) and covariance matrix is \( \Sigma_j^m \). In this text, superscript \( c, l \) or \( p \) indicates respectively that the associated expression is in the cepstral, log-spectral or power spectral domain.

Since our similarity measure is to be explicitly defined in the log-spectral domain, we will generate a set of representative vectors for each HMM state in the log-spectral domain which, when transformed to the cepstral domain, has almost the same distribution as the emitting pdf of the state. Suppose there is such a set of vectors for state \( s_j \), \( \lambda_j = \{ S_j(\omega) \} \), \( n = 1, \ldots, N \), where \( N \) is the size of the set. Then we define a similarity score, \( \rho_j \), between a test log-spectrum \( \mathcal{O}(\omega) \) and the set \( \lambda_j \), by the 1-nearest-neighbor decision rule, as

\[
\rho_j\{\mathcal{O}(\omega), \lambda_j\} = \max_{1 \leq n \leq N} \{ \rho_{jn}\{\mathcal{O}(\omega), S_j(\omega)\} \},
\]

where \( \rho_{jn} \) is the RBG similarity measure to be defined in section 3.2. During the Viterbi decoding, the similarity score \( \rho_j\{\mathcal{O}(\omega), \lambda_j\} \)
is used in place of the emitting probability defined in Eq. (1) for recognition.

To obtain the representative vectors \( \{ \tilde{S}_j(\omega) \} \) for state \( s_j \), one simple way is via the Monte-Carlo generation method [3]. First, for state \( s_j \) whose emitting pdf is defined in Eq. (1), \( N \) cepstral vectors are generated by the Monte-Carlo method. The number of the cepstral vectors generated from each mixture is proportional to the weight of that mixture. When \( N \) is sufficiently large, distribution of the generated vectors preserves the distribution defined by Eq. (1). Each cepstral vector is then transformed to log-spectral domain by an Inverse Discrete Cosine Transform (IDCT), forming a \( K \)-dimensional log-spectrum \( S(\omega) \), \( k = 1, \ldots, K \). The corresponding power spectrum, \( S'(\omega) \), is obtained by applying the exponential operation on \( S(\omega) \).

Since the RBG measure requires formant positions of \( S'(\omega) \), the autocorrelation function associated with the power spectrum \( S'(\omega) \) is first estimated by applying an inverse discrete Fourier transform on \( S'(\omega) \). Then, taking the autocorrelation function coefficients as input, the Durbin algorithm is used to estimate an all-pole model whose power spectrum optimally approximates \( S'(\omega) \). Poles of the all-pole model can be solved by the QR method [1]. Each pole \( z_i \) can be expressed in terms of its radius \( r_i \) and its angle \( \theta_i \), \( z_i = r_i \cdot e^{i\theta_i} \). This poles determination procedure will also be used in section 3.1.

For each of the poles whose angles fall into the valid frequency range of speech formants, we denote its absolute amplitude in the power spectrum \( S'(\omega) \) as \( A_i \). We then select \( Q \) poles whose \( A_i \)’s are the largest. Angles of the \( Q \) poles are sorted and stored into a vector \( P_i(k) \), \( k = 1, \ldots, Q \). Here, \( P_i(k) \) is the position of the \( k \)-th formant of \( S'(\omega) \).

For each selected pole \( z_i = r_i \cdot e^{i\theta_i} \), its partial power spectrum, \( F_{z_i}(\omega) \), is calculated by

\[
F_{z_i}(\omega) = \begin{cases} \frac{g}{(1-r_i^2)e^{-j\theta_i}} & \omega \in [\omega_i - R, \omega_i + R] \\ 0 & \text{otherwise} \end{cases}
\]

where \( z_i^* \) is the conjugate of \( z_i \), \( R = 0.17 \), and \( g \) is selected so that \( F_{z_i}(\omega_i) = A_i \). A frequency weighting vector, \( F(\omega) \), is then constructed by adding up all \( F_{z_i}(\omega) \), properly averaging the sum spectrum and transforming to the log-spectral domain. In particular,

\[
F(\omega) = \max \left\{ \sum_{i=1}^{Q} \frac{F_{z_i}(\omega)}{\sum_{i=1}^{Q} \max \{ \sum_{i=1}^{Q} |F_{z_i}(\omega)|, 1 \} } \right\}
\]

where \( \text{sign}(x) = 1 \) for \( x > 0 \) and equal to 0 otherwise.

Finally, for the \( n^{th} \) cepstral vector generated from state \( s_j \), the derived log-spectrum \( S(\omega) \), the formant positions vector \( P_i(k) \) and the frequency weighting vector \( F(\omega) \) are packed together to form a template \( \lambda_{nj} = [S(\omega), P_i(k), F(\omega)] \). A template set \( \lambda_j \) for state \( s_j \) is then formed by packing together all the \( N \) templates, \( \lambda_{nj}, n = 1, \ldots, N \). In text below, we drop the superscript \( t \) for simplicity.

As an example, Figure 1 shows such a template generated from state \( s_2 \) of triphone model [s-an+k], which denotes the Pinyin finals [an] with left context [s] and right context [k]. Regions where \( F(\omega) \) has non-zero values are the reliable bands of \( S(\omega) \). Formant positions of \( S(\omega) \) are labeled by the vector \( P_i(k) \).

3. THE RELIABLE BANDS GUIDED SIMILARITY MEASURE

3.1. Reliable Bands of Test Log-Spectrum

Under noisy conditions, it has been shown that spectral dynamic of the speech signal are quite robust to noises. Recognition performance can be improved if the spectral dynamics are properly used. In this work, spectral dynamics are used to determine which spectral bands of the test signal are more reliable than the others. Specifically, we first calculate a min-vector \( O_{min}(\omega,k) \), in the log-spectral domain for a test utterance, \( (\omega, t) \), \( t = 1, \ldots, T \), by

\[
O_{min}(\omega,k) = \min_{1 \leq \tau \leq T} \{ O(\omega,\tau) \}, \quad k = 1, \ldots, K
\]

where \( \omega_i \) is the \( k \)-th frequency band of the total \( K \) bands.

For the test log-spectrum \( O(\omega,t) \), the “poles determination” procedure outlined in section 2 is used to calculate a set of poles associated with \( O(\omega,t) \). Suppose there are \( Q \) poles whose angles fall into the valid frequency range of speech formants. We calculate the relative amplitude \( \tilde{A}_i \) of the \( k \)-th pole \( z_i = r_i \cdot e^{i\theta_i} \), by

\[
\tilde{A}_i = \exp\{O(\omega,\tau) - O_{min}(\omega_i)\}
\]

where \( \omega_i \) is the angle of the \( k \)-th pole. A score is assigned to \( z_i \) by

\[
\text{Score}(z_i) = \max_{1 \leq i \leq Q} \left\{ \tilde{A}_i \cdot \beta_{k,m} + r_i \cdot \beta_{i,t} \right\}
\]

in which \( \beta_{k,m}, \beta_{i,t} \) are two weighting factors. Determined by experiment, their optimal values are \( \beta_{k,m} = 2.0 \) and \( \beta_{i,t} = 1.0 \). Among the \( Q \) poles, we then select \( Q \) poles whose scores are the

![Figure 1: A template generated from state s2 of the triphone [s-an+k]. The Mel-Frequency Cepstral Coefficients (MFCC) representation is used to train the triphone model. 26 frequency channels are used in the calculation of the MFCC representation. Thus, the generated log-spectrum S(\omega) and the weighting vector F(\omega) both have 26 elements.](image-url)
largest. The value of $Q_k$ is determined by $Q_k = \min_j(Q_j, Q_{2j})$, where $Q_{2j}$ is a predefined constant. The frequency values of the selected poles are sorted and stored into a vector: $P_i(k), k = 1, \ldots, Q_i$.

From Eq. (6) and Eq. (7), poles introduced by slowly varying noise have low scores due to their low $A_i$’s. Contrarily, poles introduced by speech formants often have high scores due to their high $A_i$’s caused by frequently occurred formant onsets and offsets. Thus, presence of a pole in vector $P_i(k)$ often indicates that the frequency region around it is more reliable than the other bands. Consequently, vector $P_i(k)$ will be incorporated into the definition of the RBG measure in section 3.2.

### 3.2. Definition of the RBG Similarity Measure

As mentioned in section 2, for a HMM state $s_j$, a template set $\lambda_j$ which consists of $N$ templates, $\lambda_{jn}, n = 1, \ldots, N$, is generated. Each template includes one log-power spectrum $S_\omega(\omega)$, a formant positions vector $P_i(k)$, and a frequency weighting vector $F(\omega)$. The similarity score between an input frame, $\hat{S}(\omega)$, and the state $s_j$, is defined as the highest similarity score between $[O(\omega), P_i(k)]$ and all templates, $\lambda_{jn}, n = 1, \ldots, N$. In this section, we describe the definition of the RBG similarity measure between $[O(\omega), P_i(k)]$ and $\lambda_{jn}$. To define the RBG similarity measure, we first define several auxiliary functions.

Given the input log-spectrum $O(\omega)$ and the reference log-spectrum $S_\omega(\omega)$, we estimate the log-spectrum of the embedded clean speech as $\hat{S}(\omega)$, by

$$\hat{S}(\omega) = S(\omega) + \min_{0 \leq \omega \leq \pi} \{O(\omega) - S(\omega)\}. \quad (8)$$

That is, $\hat{S}(\omega)$ is obtained by “raising” $S(\omega)$ as much as possible, as long as its envelope does not exceed the envelope of $O(\omega)$.

Since the frequency weighting vectors of different templates have different overall energy levels, we normalize their overall energy levels by

$$\hat{F}(\omega) = F(\omega) \cdot \frac{C_{f,w}}{\max_{0 \leq \omega \leq \pi} \{F(\omega)\}}, \quad (9)$$

where $C_{f,w}$ is a constant. Eq. (9) is to scale each weighting vector so that its maximum amplitude reaches at the constant $C_{f,w}$.

Since speech formants are assumed to be more robust under noisy condition, we expect that the formants of the reference log-spectrum $\hat{S}(\omega)$ should closely match the formants of the input log-spectrum $O(\omega)$. Therefore, if $\hat{S}(\omega)$ has a formant within $[\omega_{s,\text{fit}}, \omega_{s,\text{left}}]$, we expect that: (a) within $[\omega_{p,\text{fit}}, \omega_{p,\text{left}}]$, $O(\omega)$ has exactly one formant whose center frequency is near that of $\hat{S}(\omega)$; (b) this formant has similar spectral shape to that of $\hat{S}(\omega)$.

To ensure (a), we require that, for any $\omega \in [\omega_{p,\text{fit}}, \omega_{p,\text{left}}]$, the nearest formant to $\omega$ in $O(\omega)$ has a center frequency which is close to the center frequency of the formant of $\hat{S}(\omega)$. The center frequency of the nearest formant to $\omega$ in the reference log-spectrum $\hat{S}(\omega)$ can be calculated by

$$P_{\text{refit}}(\omega) = \arg \min_{P_i(k)} \{ |P_i(k) - \omega| \}. \quad (10)$$

Similarly, the center frequency of the nearest formant to $\omega$ in the test log-spectrum $O(\omega)$ can be calculated by

$$P_{\text{refit}}(\omega) = \arg \min_{P_i(k)} \{ |P_i(k) - \omega| \}. \quad (11)$$

Similarity between the two center frequencies is modeled by another auxiliary function

$$f_P(x) = \max\left(x - \frac{1}{1 - b}, 0\right) \quad (12)$$

where $x$ is the difference between the two center frequency values, $\theta$ and $\omega$, are two constants whose values are chosen as $\theta = 10.0$ and $\omega = 0.3$. If $S(\omega)$ expects a formant at $\omega_i$, but $O(\omega)$ does not have a pole in the region $[\omega_i - \omega, \omega_i + \omega]$, a negative score will be produced for serious penalty. Otherwise, if $O(\omega)$ has a pole within the range, penalty will become larger as $|P_{\text{refit}}(\omega_i) - P_{\text{refit}}(\omega)|$ becomes larger.

To ensure the above (b), we require that, for any $\omega \in [\omega_{p,\text{fit}}, \omega_{p,\text{left}}]$, the amplitude difference between $O(\omega)$ and $\hat{S}(\omega)$ is small. Therefore, we define another auxiliary function to model the similarity between shapes of two formants. The function is defined by

$$f_F(x) = \frac{1}{e^{\beta(x - x_0)} - 1} \quad (14)$$

$$x_0 = \frac{\alpha \log \beta}{\beta}, \quad (15)$$

in which $\alpha$ and $\beta$ are two constants whose values are chosen as $\alpha = 1.0$ and $\beta = 0.8$.

If we want to emphasize that $O(\omega)$ should closely fit $\hat{S}(\omega)$ in the reliable bands, $\beta$ should be large. When dynamic range of the $S(\omega)$ and the $O(\omega)$ is large (possibly due to large input gain), $\alpha$ should be large to accommodate them.

Since the reliable bands of $S(\omega)$ contain formants of $S(\omega)$ and are more robust under noisy condition, they should be given more weight. Contrarily, since the unreliable bands contain spectral valleys and are more easily to be corrupted by noise, they should be given less weight. So we define another auxiliary weighting function based on the above auxiliary functions

$$W(\omega) = \left\{ \begin{array}{ll} \hat{F}(\omega) \cdot f_F(O(\omega) - \hat{S}(\omega)) \cdot f_P(1) & F(\omega) > 0 \\ 1, & F(\omega) = 0 \end{array} \right. \quad (16)$$

Finally, the RBG similarity score between $[O(\omega), P_i(k)]$ and the $j^{th}$ template of HMM state $s_j$, $\lambda_{jn} = [O(\omega), P_i(k), F(\omega)]$, is defined by

$$\rho_{jn} = \int_0^\pi \rho(\omega) \ d\omega, \quad (17)$$

where the similarity function $\rho(\omega)$ is defined by

$$\rho(\omega) = \hat{S}(\omega) \cdot W(\omega). \quad (18)$$

Figure 2 depicts the calculation procedure of the RBG measure. Among all templates generated from all model states, the template shown in the figure produces the highest similarity score. So in
4. EXPERIMENTAL RESULTS

A speaker-independent and continuously spoken Mandarin digit string recognition task is used as test-bed to evaluate the performance of the RBG measure and the PMC method. The training database consists of 6000 sentences spoken by 44 males and 31 females. The decision-tree based state-tying method is used to train a set of cross-word triphone HMMs. Each HMM has four mixtures and three states. The MFCC representation is used for the HMMs.

The clean test database consists of 1360 sentences spoken by 11 males and 6 females who are not involved in the training database. To construct noisy test database, 15 kinds of noises are drawn from the NOISEX92 database and artificially mixed with the test speech database at various SNR levels. In the NOISEX92 database, the noises are termed babble, buccaneer1, buccaneer2, destroyer, destroyerops, f16, factory1, factory2, hifchannel, leopard, m109, machinegun, pink, volvo and white.

The Data-driven Parallel Model Compensation (DPMC) algorithm is implemented on the HTK platform for comparison with the RBG measure method. For each kind of noise, a one-state HMM is trained from the corresponding noise file. The delta and acceleration coefficients are incorporated. In the implementation of the RBG measure, the state duration modelling method reported in [2] is incorporated.

Figure 3 shows the average word accuracy of the two methods as well as the baseline system. Here, the baseline system refers to using the triphone models trained under clean condition to recognize the noisy database directly. The RBG measure under-perform the DPMC method above 0 dB by 4.22% on average. This is mainly due to that the RBG measure has not yet made full use of the delta and acceleration coefficients as the DPMC does. However, it outperforms the latter by 8.8% at 0 dB. Also, its overall performance is more steady against variation of SNR levels than the latter. Moreover, it does not rely on accurate background noise modeling, which is a difficult task in itself.

5. REFERENCES


