Augmented State Space Acoustic Decoding for Modeling Local Variability in Speech

Antonio Miguel, Eduardo Lleida, Richard Rose*, Luis Buera and Alfonso Ortega

Communication Technologies Group (GTC) I3A, University of Zaragoza, Spain
{antoniomiguel, lleida, lbueera, ortega}@unizar.es

* Department of Electrical and Computer Engineering
McGill University, Montreal, Canada
rose@ece.mcgill.ca

Abstract
This paper presents a decoding method for automatic speech recognition (ASR) that reduces the impact of local spectral and temporal variabilities on ASR performance. The procedure involves augmenting the standard Viterbi search for an optimum state sequence with a locally constrained search for optimum degrees of spectral warping or temporal warping applied to individual analysis frames. It is argued in the paper that this represents an efficient and effective method for compensating for local variability in speech which may have potential application to a broader array of speech transformations. The techniques are presented in the context of existing methods for frequency warping based speaker normalization and existing methods for computation of dynamic features for ASR. The modified decoding algorithms were evaluated in both clean and noisy task domains using subsets of the Aurora 2 and Aurora 3 Speech Corpora under clean and noisy conditions. It was found that, under clean conditions on the Spanish Language Subset of the SpeechDat-Car database, the modified decoding method applied with local frequency transformations reduced word error rate (WER) by 24 percent. This was a factor of two greater reduction in WER than was obtained on the same task using the more well known frequency warping based vocal tract length normalization (VTLN) procedure.

1. Introduction
A search algorithm is presented which attempts to over come the limitations of existing methods for compensating for spectral and temporal variability in ASR. It will be referred to here as the augMented stAte space acousTic dEcoder (MATE). Vocal tract length normalization (VTLN) is a well known method used for spectral warping which has been shown to be effective in compensating for long term average mismatch between the location of spectral peaks for a test speaker and the average spectral characteristics observed in system training. These average spectral characteristics can be difficult to characterize since training conditions are represented by the statistical hidden Markov model (HMM) trained using utterances from a large population of speakers. A number of different approaches have also been proposed for the robust computation of the dynamic cepstrum in an effort to better model the dynamic temporal characteristics in speech. These temporal characteristics are often described by way of the average modulation spectrum. However, neither of these different classes of approaches attempt to represent local spectral or temporal variability in speech.

This work has been supported by MCyT under contract TIC2002-04103-C03-01

2. ML Based Spectral Warping
Frequency warping based speaker normalization techniques, often referred to as vocal tract length normalization (VTLN), have been applied in many ASR task domains [1, 2]. This class of techniques produces a warped frequency scale, \( f' = g^\circ(f) \) by selecting an optimum warping function, \( g^\circ \) from an ensemble of linear frequency warping functions, \( G = \{ g^\circ_i \}_{i=1}^\infty \). The warping functions are of the form illustrated by the curves in Figure 1. The optimum warping function,
The sampling frequency.

The procedure for obtaining the frequency warped cepstrum, \( C^\alpha \), involves modifying the Mel-frequency filter bank that is used in Mel-frequency cepstrum coefficient (MFCC) feature analysis [2]. The \( B \) channel filter bank is implemented as a set of \( B \) triangular weighting functions applied to \( S \) discrete spectral magnitude values. The entire set of \( S \) component spectral magnitude values for an \( L \) frame utterance can be represented as an \( S \times L \) dimensional matrix, \( X \). The \( D \) component cepstrum vectors are computed as the discrete cosine transform of the log of the outputs of these filters. Frequency warping can be implemented by applying the warping functions, \( g^\alpha()\), to the array of filter bank coefficients which can be represented as a \( B \times S \) dimensional matrix, \( F^\alpha \). The set of warped cepstrum vectors for an \( L \) frame utterance can then be expressed in vector notation as

\[
C^\alpha = DCT \cdot \log(F^\alpha \cdot X),
\]

where \( DCT \) is a \( D \times B \) matrix of cosine coefficients representing the discrete cosine transform applied to the filter bank outputs. This same method for producing warped cepstrum will also be used in the augmented state space decoder described in Section 3.

While this maximum likelihood based warping procedure has been shown to significantly reduce WER in many cases, it has two important limitations. The first is that it can be unwieldy to apply. It is generally implemented as a two pass procedure which can make real-time implementation difficult. The first pass is used to generate an initial hypothesized word string. This initial word string is then used in a second pass to compute the likelihood of each warped utterance by aligning \( C^\alpha \) with the decoded word string.

The second limitation is related to the fact that only a single linear warping function is selected for an entire utterance. Even though physiological evidence indicates that all phonetic events do not exhibit similar spectral variation as a result of physiological differences on vocal tract shape, this technique estimates a single transformation for an entire utterance. The procedure described in Section 3 addresses both of these issues. The procedure requires only a single pass over the input utterance and produces frame-specific estimates of the frequency warping functions.

### 3. Augmented state space decoder

This section presents a Viterbi algorithm that is implemented in an augmented state space. It allows frame-specific spectral warping functions to be estimated as part of search. It will also be shown in Section 4 that this same search procedure can be used to estimate the optimum frame-specific time interval over which cepstrum difference coefficients are computed. The section begins with a description of the augmented state space, and then the modified search algorithm is presented.

A Viterbi beam search decoder for continuous speech recognition is implemented by propagating paths into the nodes of a two dimensional trellis. Each node of the trellis corresponds to one of \( M \) HMM states \( \{q_t\}_{t=1}^M \) evaluated for observation vectors \( e_t, t=1,...,L \). In the MATE decoder, the state space can potentially be expanded by a factor of \( N \), where \( N = N_o \) is the size of the warping function ensemble described in Section 2. This effectively results in a three dimensional trellis. Each node of this augmented trellis now corresponds to one of as many as \( M' = N \cdot M \) states, \( \{q_t\}_{t=1}^{M,N} \).

In the current implementation, the states \( \{q_t\}_{t=1}^{M,N} \) share the same observation densities as the state \( q_j \) in the original model for all \( j = 1,...,M \). This tying of the observation densities can be expressed as

\[
b_j^k(c_i) = b_j(c_i), \quad j = 1,...,M, \quad k = 1,...,N.
\]

In Equation 3, \( \phi_j(t) \) is the likelihood of the optimum path terminating in HMM state \( q_j \) at time \( t \) and \( a_{i,j} \) is the transition probability from state \( q_i \) to state \( q_j \). The \( max \) is computed over all states that are permitted by the HMM model to propagate into state \( q_j \) which, for a left-to-right HMM topology would be \( q_{j-1} \).

In the MATE decoder, the optimum sequence of states is identified for the decoding process in a standard HMM using the Viterbi algorithm,

\[
\phi_j(t) = \max_i \{ \phi_i(t-1) \cdot a_{i,j} \} \cdot b_j(c_i).
\]

In Equation 4, \( \phi_{j,n}(t) \) is the likelihood of the optimum path terminating in state \( q_j^m \) at time \( t \) and \( a_{i,j}^m \) is the transition probability from state \( q_j^m \) to state \( q_i^m \). The \( max \) is computed over all states that are permitted by the HMM model to propagate into state \( q_j^m \).

Just as structural constraints can be placed on standard HMM topologies by constraining the allowable transitions between HMM states, constraints can also be placed on the transformations, \( g^\alpha \), that are permitted at state \( q_j^m \) in the augmented state space. These constraints can be applied by setting a subset of the transition probabilities, \( a_{i,j}^m \), equal to zero. In this work, transition probabilities were constrained so that the frequency warping transformations applied to adjacent frames were required to be taken from adjacent indices in the ensemble \( G \).

\[
a_{i,j}^m = 0, \text{ if } |m-n| > 1.
\]

These constraints have the effect of reducing the computational complexity in search. Furthermore, they also provide a means for limiting the degrees of freedom in the application of spectral transformations to reflect a more physiologically plausible degree of variability. Additional constraints can be applied. For example, HMM states for non-speech models are constrained so that no transformations can be applied to the observation vectors that are associated with those states.
An illustration of the effect of the MATE decoder is provided by the spectrogram and log likelihood plots for an utterance of the word “two” shown in Figures 2a and 2b respectively. Figure 2b displays the log likelihoods plotted versus time for a set of \( N_v = 20 \) possible frequency warping transformations that correspond to compressing and expanding the frequency axis by as much as fifteen percent. This plot was obtained by performing Viterbi decoding on this utterance using the modified Viterbi algorithm given in Equation 4. The trajectory of warping functions corresponding to the optimal path is indicated by a superimposed white line. It is interesting that, for the initial part of the word corresponding to the unvoiced aspirated “t” sound, values of \( \alpha \approx 1.0 \) are chosen. This effectively corresponds to the choice of no spectral transformation for this region. However, in the voiced part of the word, a warping value is chosen that is similar to the value chosen by the global linear VTLN warping factor that had been estimated for this utterance (\( \alpha = 1.13 \)).

We discovered previously published work directed at extending the search space in ASR by estimating local frequency warping parameters [3]. While there were differences in the manner in which the search procedure was performed and the manner in which constraints were applied, the general motivation when applied specifically to frequency warping was similar to the motivation for the work presented here. The next section, however, presents an example of one of potentially similar motivation when applied specifically to frequency warping was introduced in Section 3 can be applied to selecting the frame update interval so that a modified window of cepstrum frames centered at frame index \( t \) used for computing the dynamic cepstrum with a temporal resolution of \( \beta \) is shown in Figure 3 for \( M = 2 \) as \( C_w(t, \beta) \).

Given an ensemble of resolution factors \( \mathcal{B} = \{ \beta_i \}_{i=1}^{N_{\beta}} \), the dynamic features can be locally optimized using the modified Viterbi algorithm in Equation (4). The frequency warped parameters \( c_1 \) are replaced by the new \( \beta \) dependent parameters, \( c_1^{\beta_i} \). The same local constraints that were applied in Section 3 to limit the rate of change of the frequency warping parameters will also be applied to the temporal resolution parameters.

It is clear from Figure 3 that the above procedure requires that MFCC cepstrum vectors be computed at every sample instant corresponding to the ensemble of resolution factors in the ensemble \( \mathcal{B} \). To account for this, a vector sequence \( C_1 \) is defined that contains \( D \) dimensional cepstrum vectors computed for each sample. The original feature vector sequence \( C = \{ c_1 \}_{t=1}^{T_{\text{original}}} \) represents a subsampled version of \( C_1 \), where \( c_1 = c_{1(tT)} \). The window of cepstrum frames centered at frame \( t \) can be written in terms of the elements of \( C_1 \) as

\[
C_w(t) = \left[ c_{1(tT-\Delta T)}, \ldots, c_{1(tT)}, \ldots, c_{1(tT+\Delta T)} \right].
\]  

The feature vector corresponding to the concatenation of the static cepstrum and the first and second order difference cepstrum is calculated as the linear combination of the static feature vectors over \( C_w(t) \).

The resolution factor, \( \beta \), is used to modify the manner in which the dynamic cepstrum is computed by modifying the frame update interval so that a modified window of cepstrum frames, \( C_w(t, \beta) \), is obtained. The new frame update interval \( T' \), can be expressed as \( T' = \beta \cdot T \) and the locally warped window of cepstrum frames can be expressed as:

\[
C_w(t, \beta) = \left[ c_{1(tT' \cdot \beta T)}, \ldots, c_{1(tT')}, \ldots, c_{1(tT'+\Delta T)} \right].
\]

By estimating \( \beta \) using the augmented state space decoding algorithm in Section 3, frame specific temporal resolution is obtained for dynamic features that maximizes the likelihood of the utterance with respect to the original HMM. Section 5 will present an experimental study that demonstrates the impact of this temporal normalization procedure on ASR WER.
5. Results

An experimental study was performed to evaluate the use of the MATE decoder in performing local optimization of both spectral warping features as discussed in Section 3 and temporal warping features as discussed in Section 4. The task domain was based on the Spanish Language subset of the Speech-Dat-Car database used for the Aurora 3 task and the “A” subset of the Aurora 2 task that includes subway, babble, car, and exhibition noises.

HMM word models with 16 states and 3 Gaussians per state were used to model the vocabulary of spoken digits. Initial and final silence were modeled by 3 state HMMs with 6 Gaussians per state. Inter-word silence was modeled by 1 state HMM with 6 Gaussians. The parameters used in the experiments were the standard and the advanced ETSI front end, both with a window size of 25 msec. and a 10 msec. frame update interval. Twelve cepstrum coefficients, energy, velocity, and acceleration parameters were computed for each frame resulting in a total of 39 parameters. The baseline models were obtained with 20 training iterations and used to build MATE models according to the observation densities given in Equation 4. Retrained MATE models were obtained by using the modified Viterbi alignment given in Equation 4.

Table 1 shows the experimental results obtained using noise free speech in order to compare performance of the local temporal and frequency optimization techniques to the original baseline performance using the standard ETSI front end. The experiments were performed on two data sets. First, the Aurora2 clean training set was used which consists of 8440 phrases (27727 digits) and test set consists of 4004 phrases (13159 digits). Second, the Spanish language subset of the Aurora2, Speech-Dat-Car corpus. This improvement may be a result of the greater specificity of the speech compared to previous VTLN methods. MATE-frequency gives more than a factor of two in reduction of WER over that obtained using VTLN. MATE-time was shown to provide more than 15% WER reduction when used in real and stressed situations.

These results suggest that there may be other speech transformations that may be applied in the context of the constrained search algorithm described here. The MATE decoder can be seen as a method for making HMMs more “locally elastic” in providing a mechanism for better tracking of the dynamics of speech variability. Application to domains that characterize more extreme speaker variability arising from increased vocal effort or task related stress will provide a further indication of the potential benefits of these techniques.

6. Conclusions

An augmented state space acoustic decoding technique, MATE, has been presented. The technique provides a mechanism for either the spectral warping either the dynamic feature computation to be locally optimized. In the MATE decoder, the optimum sequence of states in the augmented state space is identified using a modified Viterbi algorithm. It allows frame specific spectral warping functions or temporal resolution for the dynamic features to be estimated as part of search.

Our approach has been tested on subsets of the Aurora 2 and Aurora 3 task domains under clean and noise conditions in the context of existing methods for frequency warping based speaker normalization and existing methods for computation of dynamic features. The results have shown that MATE is an effective method for compensating the local variability of the speech compared to previous VTLN methods. MATE-frequency gives more than a factor of two in reduction of WER over that obtained using VTLN. MATE-time was shown to provide more than 15% WER reduction when used in real and stressed situations.

7. References


<table>
<thead>
<tr>
<th>DataBase</th>
<th>Aurora2</th>
<th>Speech-Dat-Car</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td>VTLN</td>
<td>0.85 (+6%)</td>
<td>0.81 (+8%)</td>
</tr>
<tr>
<td>retrained-VTLN</td>
<td>0.87 (+3%)</td>
<td>0.77 (+13%)</td>
</tr>
<tr>
<td>MATE-freq</td>
<td>0.81 (+10%)</td>
<td>0.68 (+23%)</td>
</tr>
<tr>
<td>retrained-MATE-freq</td>
<td>0.78 (+13%)</td>
<td>0.67 (+24%)</td>
</tr>
<tr>
<td>MATE-time</td>
<td>0.88 (+2%)</td>
<td>0.75 (+15%)</td>
</tr>
</tbody>
</table>

Table 1: Clean tests recognition results in error rate