PROSPECT Features and their Application to Missing Data Techniques for Vocal Tract Length Normalization

Wim Jansen, Hugo Van hamme

Katholieke Universiteit Leuven
Dept. ESAT, Kasteelpark Arenberg 10, B-3001 Heverlee, Belgium
{wim.jansen,hugo.vanhamme}@esat.kuleuven.ac.be

Abstract

Speaker normalization by (piecewise) linear warping of the frequency axis is a popular method because of its simplicity and effectiveness. However, when this so-called vocal tract length normalization is applied to map test speakers with a shorter vocal tract onto acoustic models trained on speakers with a longer vocal tract, there is important information missing in the frequency bins at the high end of the spectrum. Usually, this missing information is reconstructed by ad hoc rules or through extrapolation of the spectrum. In this paper, we present a new method to estimate the content of those bins. The proposed solution is derived from Missing Data Techniques, that are used for noise robust speech recognizers. To alleviate the accuracy loss associated with Missing Data Techniques that are usually expressed in the spectral domain, we apply the PROSPECT feature representation introduced about a year ago. We demonstrate the superiority of our approach on the TIDigits database.

1. Introduction

The spectral properties of male, female and child speech differ in a number of ways (e.g. [1]). One prominent distinction is caused by the difference between their average Vocal Tract Length (VTL). In fact, the VTL of women is about 10% shorter compared to the VTL of men. The VTL of children is even shorter (up to 10%) than that of women. According to the linear acoustic theory of speech production, this directly implies that all the formants in children’s speech are subjected to a (fixed, VTL-dependent) scaling towards the high end of the spectrum. In order to create a good match with acoustic models trained on adults, one has to warp the spectra extracted from children’s speech towards the lower end. However, when the child data are sampled at the same frequency as the adult data, the required scaling of the frequency axis creates a problem at the high end of the spectrum: there is no spectral information above the Nyquist frequency. Classically, this problem is tackled by a non-linear warping, such as a piecewise linear function [2], to approximate the spectral content at the high end of the spectrum. In this paper, we will treat the unknown spectral information using Missing Data Techniques (MDT).

This paper is organized as follows. Section 2 describes VTLN and the inherent problem that is associated with it. The next section explains in general the missing data approach to solve that problem. Section 4 describes the PROSPECT features that will be used in this paper. The last section gives an overview of the experiments on the TIDigits database.

2. Vocal Tract Length Normalization (VTLN)

In Vocal Tract Length Normalization, speech spectra are warped to a different frequency axis. Signal energy at \( f_{\text{data}} \) is processed as if it occurred at \( f_{\text{model}} \). Due to the shorter VTL, the formants in children’s speech will occur at higher frequencies, hence for a model trained on adult data, we want \( f_{\text{data}} = \alpha f_{\text{model}} \) for some warping factor \( \alpha > 1 \). However, there is no data to warp onto the frequency range \( \alpha^{-1} f_{\text{Nyq}} < f_{\text{model}} \leq f_{\text{Nyq}} \). Figure 1 illustrates this warping to lower frequencies. The grey area on the vertical frequency axis does not have information, as there is no counterpart on the horizontal frequency axis. A similar effect is observed when warping to the high end of the spectra (\( \alpha < 1 \)), but now information has to be discarded. There exist many methods to solve this information problem. A popular approach is to use a piecewise linear transformation [2], which has as advantage that the Nyquist frequency is always warped to the Nyquist frequency, and therefore the complete bandwidth will be used. However, there is no physical rationale for the expansion of the frequency axis above the knee frequency (\( f_{\text{knee}} \), see figure 1). Also, the warping parameter has a relation with the VTL which correlates with age and gender, which is not the case for the knee frequency of the piecewise linear transformation. Other techniques use a non-linear warping function [3], time domain interpolation or low pass filtering extrapolation [4].

In this paper, we consider a filterbank-based front-end to
extract a smoothed spectrum from the incoming speech signal. The shape of the filters or their overlap is in principle of no importance to the proposed method. Experimental evidence will however only be provided using MEL-scaled filterbanks. In this setting, warping is implemented by re-assigning the observed energy in a Fourier bin to a different frequency according to the warping function. We maintain the linear warping \( f_{data} = \alpha f_{model} \) for \( 0 \leq f_{data} \leq f_{Nyq} \), so the filter banks at the highest center frequencies will receive incomplete or no energy contributions when \( \alpha > 1 \). Any filter bank that expects an energy contribution with \( f_{model} > f_{Nyq} \cdot \alpha^{-1} \) will be considered as generating an unreliable feature and will receive a special treatment as explained below.

3. Missing data techniques

Missing data techniques for noise robust speech recognition rely on the property that some regions in the spectrogram will be dominated by the speech signal that is to be recognized, while other regions will be dominated by unwanted signals such as noise or competing speech. A spectral mask defines where on the one hand spectral information is reliable and can be used as such, and on the other hand where it is unreliable and where an acoustic model trained on clean speech needs to be modified to assure a good match with the corrupted data. In [5], several methods for adapting continuous-density HMMs were presented.

To apply MDT to the case of VTLN, some care must be taken. First, missing data are identified in the spectral domain. In contrast to the implementations dealing with noise robustness, the mask estimation from the frequency warping function becomes trivial as explained above. Secondly, in noise robustness research, the signal energy in an analysis filter serves as an upper bound for the clean speech energy (the signal contains missing data) and will receive incomplete or no energy contributions. Even in [6], it has been shown that these features can be modeled well by a GMM with diagonal covariance, even for \( K \) as small as 3. The likelihood of the \( i \)-th mixture component of HMM state \( q \) has the expression

\[
\begin{align*}
    f(p|i, q) &= N(c|i, q) \cdot N(d|i, q)^\beta \\
    &\text{where } \beta \text{ is a stream exponent, and } N \text{ represents a multivariate Gaussian distribution with diagonal covariances.}
\end{align*}
\]

4.1. PROSPECT and VTLN

The problem of the missing information when warping a spectrum to lower frequencies can now be formulated using missing data theory. Let \( y \) be the observed filter bank spectrum after linear warping with \( f_{data} = \alpha f_{model} \) for \( 0 \leq f_{data} \leq f_{Nyq} \). Suppose a wideband version of the same signal would be available, such that the incomplete filter bank energies in \( y \) can be computed by linear warping \( f_{data} = \alpha f_{model} \) for \( 0 \leq f_{data} \leq \alpha f_{Nyq} \), and imputed. Let \( s \) be the resulting (unknown) spectrum. Then the inequality

\[
s_m \geq y_m
\]

holds, where subscript \( m \) denotes the components of \( y \) that are missing or incomplete. The \( (q-\text{and } i\text{-dependent}) \) maximum likelihood estimator of \( s \) is now obtained by maximizing log-likelihood (4) with respect to \( s_m \) subject to inequality 6. Hence, the imputed spectrum is

\[
\arg\min_{s_m} \left[ \frac{1}{2} \left( \begin{bmatrix} s_r \\ s_m \end{bmatrix} - \mu \right)^\top A \cdot \left( \begin{bmatrix} s_r \\ s_m \end{bmatrix} - \mu \right) \right]
\]

where the shorthand \( A \) was introduced for the precision matrix in 5 for ease of notation in the sequel. This results also in a NNLSQ problem (see section 3), and possible efficient iterative solutions have been discussed in [6].
4.2. MDT with normalization for convolutional distortion

In [8], a normalization method for convolutional distortion was proposed, which consists in averaging the spectral features over their \( L \) largest values. In the present case, there is however a problem. Since there is never information available in the high end of the warped spectrum, it is impossible to give an estimate of the log-spectral means at these frequencies. More sophisticated approaches to estimating the convolutional distortion take a maximum likelihood approach, i.e. they add a constant \( c \) to \( s \) and estimate \( c \) by maximizing the likelihood of an ensemble of observations. Whereas this will work fine for the complete components, it is equivalent to changing the constraint to \( s_m \geq y_m - c_m \). Hence the maximum likelihood solution will be an arbitrarily large value for the missing components \( c_m \), since this is equivalent to removing the constraint, which will always increase the likelihood. Hence in a ML formulation with convolutional noise removal, the exact solution can be obtained by leaving out the constraint. If we partition the precision matrix in four different parts (one for each combination of reliable and missing, see equation 8, where \( \mathbf{A}_{mr} = \mathbf{A}_{rm}^t \)), the minimum of the log-likelihood function can be calculated with equation 9.

\[
\mathbf{A} = \begin{bmatrix} \mathbf{A}_{ss} & \mathbf{A}_{sr} \\ \mathbf{A}_{sr}^t & \mathbf{A}_{rr} \end{bmatrix}
\]

\[
\frac{1}{2} (s_r - \mu_r)^t [\mathbf{A}_{rr} - \mathbf{A}_{mr} \mathbf{A}_{rm}^t] (s_r - \mu_r)
\]

Alternatively, like the constrained minimization in equation 7, unconstrained iterative gradient methods can be applied as well.

5. Experimental Results

The proposed approach was evaluated on the TIDigits database. This database contains digit strings from male and female adult speakers, and also from boys and girls.

The acoustic models were trained from the AURORA-2 [9] clean training data which is a subset of the TIDigits training data containing adult speakers only and which is downsampled to 8 kHz. We used the AURORA WI-007 MFCC front end and the complex back-end (16 HMM states per digit and 20 Gaussians per state) using the standard training scripts. Then models for the PROSPECT representation are derived by single pass retraining as described in [6]. The test data contain all 3847 test-utterances of boys and girls aged 6 to 15. For most tests the spectral data above \( f_{N_yq} \) (4 kHz) are discarded, but thanks to the 20 kHz sampling rate, we can also compare exploiting wideband data.

5.1. Results without Convolutional Distortion Normalization (CDN)

Figures 2 and 3 show the word error rate as a function of the warping factor. The baseline VTLN experiment uses a piecewise linear transformation with \( \alpha \) frequency at 3.5 kHz (labeled \textit{bandlimit}). The results using the CMDT (labeled \textit{cmdt}) and PROSPECT (labeled \textit{prospect}) missing data approach are also given. Obviously, they both outperform the baseline, with an advantage for the PROSPECT method. In the previous experiments, MDT was only applied to the static features. It is much more difficult to include the dynamic features in the optimization, since this problem has a much bigger search space [7].

5.2. Results with CDN

Therefore, as an intermediate solution, the next experiment (labeled \textit{prospect + delta}) uses a piecewise linear transformation for the dynamic features, and PROSPECT features for the static features. This resulted in an extra relative improvement of 22% when evaluated at the girls data (\( \alpha = 1.20 \)) with respect to the previous experiment, but only insignificant changes were noticed on the boys data.

Finally, the experiment labeled \textit{wideband} uses the linear warping relation for \( 0 \leq \alpha \leq f_{N_yq} \) and hence exploits the data spectrum up to \( \alpha f_{N_yq} \). Surprisingly, this approach scores worse than piecewise linear warping. Closer examination reveals that this can be attributed to the widening of the analysis filters caused by the warping. Applying the same filter gain as introduced by the piecewise linear warping reduces the error rate of the wideband results below those of piecewise linear warping and around those of the MDT methods (experiment labeled \textit{norm energy}). Notice that the MDT techniques do not suffer from these energy normalization problems since they adjust (estimate) the energy automatically.
strum are related by a matrix multiplication, this is equivalent to cepstral mean subtraction (CMS). Hence, in the second series of experiments (see figures 4 and 5), CDN has been applied to the features during training and testing.

The first VTLN experiment uses the piecewise linear transformation defined in section 5.1. Notice that the overall results are better with CDN (a difference of approximately 1% point). As expected, the wideband results are better than the piecewise linear warping. Again, the PROSPECT method (now without constraints) outperforms the piecewise linear warping. When only applying the piecewise linear transform to calculate the dynamic features, there is again a small improvement. On the boys data, the resulting accuracy is even better than what is observed on the wideband signal. This effect is attributed to the freedom of the MDT approach to optimize the energy values at high frequencies and hence deviate from linear warping if necessary.

6. Conclusions

We successfully applied the missing data techniques to solve the information problem that always occurs when applying VTLN to the recognition of children’s speech with acoustic models trained on adult speakers. Missing Data Techniques are normally used for noise robust speech recognition. In this paper, it has been shown that the same techniques can be applied successfully to estimate the unknown speech spectrum that is missing at high frequencies due to frequency axis warping. Essential in our approach is to use the PROSPECT feature representation, which allows high accuracy modeling in conjunction with MDT at a reasonable computational load. The traditional approach of using piecewise linear warping was inferior to all the methods that have been proposed in this paper. On the boys data of the TIDigits database, the new methods even performed better than the wideband signal, which uses a higher sampling frequency.

7. References


