NEW TRANSFORMATIONS OF CEPSTRAL PARAMETERS FOR AUTOMATIC VOCAL TRACT LENGTH NORMALIZATION IN SPEECH RECOGNITION *

Tom Claes (2), *Ioannis Dologlou* (1), *Louis ten Bosch* (2), *Dirk Van Compernolle* (2)

(1) *K.U.Leuven - E.S.A.T.*, Kardinaal Mercierlaan 94, B-3001 Heverlee, Belgium
(2) *Lernout & Hauspie Speech Products*, Koning Albert I laan 64, B-1780 Wemmel, Belgium

ABSTRACT

This paper proposes a method to transform acoustic models (HMM gaussian mixtures) that have been trained on a certain group of speakers for use on speech from a different group of speakers. Cepstral features are transformed on the basis of assumptions regarding the difference in *vocal tract length* (VTL) between the groups of speakers (VTL normalisation, VTLN). Firstly, the VTL of these groups has been estimated based on the average third formant F_3. Secondly, the linear acoustic theory of speech production has been applied to warp the spectral characteristics of the existing models so as to match the incoming speech.
The mapping is composed of subsequent non-linear submappings. By locally linearizing it, a linear approximation was obtained which is accurate as long as warping is reasonably small.
The method has been tested for the TI digits database, containing adult and kids speech, consisting of isolated digits and digit strings of different length. The word error rate when trained on adults and tested on kids with transformed adult models is decreased by more than a factor of 2 compared to the non-transformed case.

1. INTRODUCTION

This paper proposes new methods to transform automatically acoustic models that have been trained with a certain group of speakers and make possible the efficient use of these models when other speakers with different vocal tract characteristics are tested.

It is known [1][2][3] that the spectral properties of male, female and child speech differ in a number of ways. One prominent difference is due to the difference between their average vocal tract length (VTL). The VTL of females is about 10% shorter compared to the VTL of males, and the VTL of children is up to 10 % shorter than that of females. According to the linear acoustic theory of speech production (cf. [2]), this implies that, compared to male speech, all the formants in female and kids speech undergo a (fixed, VTL-dependent) scaling towards the high end of the spectrum. Consequently, one has to warp the children spectra towards the lower end in order to match it with the adult (female and male) speech. In this paper, the problem of how to do so is addressed.

A related issue to be discussed here is the estimation of the VTL from a given speech signal. The VTL-estimation is used subsequently to determine the frequency warping factor. The VTL is related to the position of formants and in particular to the position of the third formant (F_3), which is less influenced by the vowel under consideration while its detection from the signal itself remains quite reliable. For that reason we will base the estimation of both the VTL and the warping factor on F_3 estimations.

The feature vector, which is computed from a speech frame, consists of the mel scale cepstral coefficients (MFCC) (12 cepstra) along with 12 delta-cepstra, 12 delta^2-cepstra and delta-log(Energy), delta^2-log(Energy), i.e. 38 parameters in total.
The MFCC parameters [6] are calculated as follows. Based on the power spectrum, a mel scale spectrum is calculated using a simulated filterbank with triangular filters. The filter centers are linearly spaced below 1kHz and logarthmically above that. The mel cepstrum is then calculated by taking the logarithm and the inverse cosine transform (IDCT). The mel scale simulated filterbank is the only step that cannot be inverted accurately and an approximative scheme is proposed to do so.

In the following, first a non-linear transformation relating the MFCC of kids and adults is derived and it is shown how this can be approximated linearly. Thereafter, the computation of the transformed means and covariances of the hidden markov models takes place. It is shown that a *linear approximation* is accurate as long as warping is reasonably small. To evaluate the proposed methodologies experiments for continuous word speech recognition have been carried out using the digit-strings

*This research work was carried out at the ESAT laboratory of the Katholieke Universiteit Leuven, supported by the *KIDS* Research Contract nr. ITA/950226 from the IWT.

of the TI-digits database (available through the Linguistic Data Consortium). Tests are performed using continuous density HMM's with one model per word (11 states per word).

2. WARPING TECHNIQUE FOR VOCAL TRACT LENGTH NORMALISATION

The difference in average VTL for male, female and kids is an important problem for speech recognition with children compared to speech recognition with adults. For average formant frequencies, we have [1]

$$\text{VTL} \approx \frac{(2i-1)c}{4\text{F}_i} \tag{1}$$

where c is the velocity of the sound. So

$$\frac{\text{VTL}_{adults}}{\text{VTL}_{kids}} \approx \frac{\text{F}_{i,kids}}{\text{F}_{i,adults}} \tag{2}$$

Therefore normalisation of the VTL can be achieved by a (linear) frequency warping procedure.

According to [5], MFCC coefficients [6] give better recognition performance with children than LPC's, because of the mel scale frequency warping. Many authors indicate that the use of frequency warping procedures improves Speaker Independent speech recognition performance with males and females [4] . This paper combines the concept of frequency warping and the use of MFCC's [4] and translates them into a transformation of the cepstra.

2.1. The warping factor wp

The main problem with the frequency warping techniques is to determine the warping factor as a function of the VTL. We will follow [4][1] and use the (averaged) third formant F_3. Presently, the warping factor wp is defined as the ratio of the median of F_3 of the speaker test class to the median of F_3 of the training class.

The following algorithm is proposed for the calculation of the transformed speech parameters.

1. estimate F_3 for all frames.
2. calculate the median of F_3.
3. warp the frequency axis based on the ratio of the calculated median to the median of the training data.
4. calculate the speech parameters based on the warped spectrum

3. MEL CEPSTRUM TRANSFORMATION

Let **F** denote the matrix to be applied to the FFT power spectrum in order to obtain the mel-scaled spectrum. **F** is not invertible, but has a pseudo-inverse. The following approach is used to deduce the spectrum on a linear frequency scale that provides a better approximation after warping and filterbank.
For $i = 1, \ldots, n_c$

- let S_i be the energy of output-channel i of the filterbank
- set equal to S_i the fft-point which corresponds to the center of channel i
- set the other frequency points of the fft equal to zero

This algorithm may be represented by the multiplication of S with matrix **A** such that **F.A=I**. Note that the product **A**.S provides the power spectrum and warping consists in simple multiplication with **W**. As a result the complete transformation for the mel-spectrum S, can be expressed using the following matrix **T** :

$$\mathbf{T} = \mathbf{F} \,.\, \mathbf{W} \,.\, \mathbf{A}$$

where the matrix **W** denotes the warping matrix. $P_2 = \mathbf{W}.P_1$ is the linearly warped power spectrum given the original spectrum P_1.

To deduce the transformation of the mel cepstrum c (MFCC parameter), let **C** be the matrix representing the DCT transform, the mel-spectrum S corresponding to c is given by $S = \exp\left(\mathbf{C}^{-1}.c\right)$. A simple multiplication of S with matrix **T** and the computation of the logarithm and the DCT-transform of the result provides the transformed equivalent in the cepstral domain. This can also be expressed by the following relationship between the initial MFCC parameter set c and the transformed MFCC parameter set $\tilde{c}$:

$$\tilde{c} \quad = \quad \mathbf{C}.\log\left\{\mathbf{T}.\exp\left(\mathbf{C}^{-1}.c\right)\right\} \tag{3}$$

3.1. Theoretical derivation of the transformed model parameters

Transformation of the static model parameters The transformation of a single cepstral vector (static mel cepstra) is calculated using equation 3. Given the means of the cepstra μ_c and the covariance matrix Σ_c, the means $\tilde{\mu}_c$ and the covariance matrix $\tilde{\Sigma}_c$ of the transformed cepstra is required. This is not a trivial problem because the transformation is not linear and the following holds for the expected values :

$$\begin{aligned} E\{\tilde{c}\} &= E\left\{\mathbf{C}.\log\left\{\mathbf{T}.\exp\left(\mathbf{C}^{-1}.c\right)\right\}\right\} \\ &\neq \mathbf{C}.\log\left\{\mathbf{T}.\exp\left(\mathbf{C}^{-1}.E\{c\}\right)\right\} \end{aligned}$$

Thus, the transformed model cannot be expressed by the following mean vector

$$\tilde{\mu}_c \quad \neq \quad \mathbf{C}.\log\left\{\mathbf{T}.\exp\left(\mathbf{C}^{-1}.\mu_c\right)\right\}$$

This would be the case if the transformation was linear. In [8], the expressions for $\Delta\tilde{c}^{(i)}$, $E\left\{\Delta\tilde{c}^{(i)}\right\}$, and $E\left\{\Delta^2\tilde{c}^{(i)}\right\}$ has been explicitly derived by using formula 3.

3.2. Linear approximation of the transformation

By applying $\log(1+x) \approx x$ and $\exp(x) \approx 1+x$ for matrices, it is straightforward to derive a linear approximation for equation 3. It can be shown ([8]) that transformation 3 may be adequately approximated by a linear one of the form,

$$\tilde{c} \quad \approx \quad a + \mathbf{B}.c \tag{4}$$

where

$$\begin{aligned} a &= \mathbf{C}.\begin{bmatrix} \log\left\{\sum_j t_{1,j}\right\} \\ \cdots \\ \log\left\{\sum_j t_{n_c,j}\right\} \end{bmatrix} \\ \mathbf{B} &= \mathbf{C}.\begin{bmatrix} \frac{1}{\sum_j t_{1,j}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sum_j t_{n_c,j}} \end{bmatrix}.\mathbf{T}.\mathbf{C}^{-1} \end{aligned}$$

and $t_{i,j}$ are the elements of transformation matrix **T**.

Transformation of the static and dynamic model parameters based on the linear approximation When the linear approximation is applied to the mel cepstrum the statistics of the transformed models can be calculated as ([8]) :

$$\begin{aligned} E\{\tilde{c}\} &\approx a + \mathbf{B}.E\{c\} \\ \tilde{\mu}_c &\approx a + \mathbf{B}.\mu_c \\ &\approx \mathbf{C}.\log\left\{\mathbf{T}.\exp\left(\mathbf{C}^{-1}.\mu_c\right)\right\} \end{aligned}$$

and

$$\tilde{\Sigma}_c \quad \approx \quad \mathbf{B}.\Sigma_c.\mathbf{B}^t \tag{5}$$

Likewise, we use the approximations

$$\begin{aligned} \Delta\tilde{c} &\approx \mathbf{B}.\Delta c \\ \Delta^2\tilde{c} &\approx \mathbf{B}.\Delta^2 c \end{aligned}$$

which can be applied directly to the mean values,

$$\begin{aligned} \tilde{\mu}_{\Delta c} &\approx \mathbf{B}.\mu_{\Delta c} \\ \tilde{\mu}_{\Delta^2 c} &\approx \mathbf{B}.\mu_{\Delta^2 c} \end{aligned}$$

The covariances are transformed as :

$$\tilde{\Sigma}_{\Delta c} \quad \approx \quad \mathbf{B}.\Sigma_{\Delta c}.\mathbf{B}^t \tag{6}$$

$$\tilde{\Sigma}_{\Delta^2 c} \quad \approx \quad \mathbf{B}.\Sigma_{\Delta^2 c}.\mathbf{B}^t \tag{7}$$

The accuracy of this approximation is assessed in [8].

4. EXPERIMENTS

In this section the derived transformations are used for speech recognition with kids. The goal is to use models that are trained with adult speech for the recognition of kids.
All experiments have been carried out using the TI-digits database.

- sampling frequency 16 kHz (originally at 20 kHz)
- isolated digits and digit strings of different length
- 22 isolated digits and 55 strings per speaker
- training : 55 men + 57 women or 25 boys + 26 girls
- testing : 25 boys + 25 girls

4.1. Calculation of the warping factor

The warping factor related to two speaker classes is defined as the ratio of the mean F_3 of the classes. The average F_3 and corresponding VTL for each class are given in table 1.

class	mean of median F_3 (Hz) $\pm\,\sigma$	VTL (cm)
M	2497 ± 363	17
F	2813 ± 349	15
K	3071 ± 427	14

Table 1: Average third formants of males, females and kids and corresponding VTL

The VTL results correspond very well to the results given in [7]. This indicates that the mean third formant (F_3) is closely related to the VTL. We get:

$$wp = \frac{F_{3,kids}}{F_{3,adults}} = 1.2 \tag{8}$$

with $F_{3,kids}$ and $F_{3,adults}$ the average F_3 for the kids and adults class respectively.

4.2. Recognition Experiments

All tests were performed on continuous word recognition using Continuous Density HMM's with 1 model per word (11 states). The training data of the adults (males and females) is used for training the Hidden Markov Models. The generated models are represented by their μ and Σ. In general only the diagonal elements of Σ are known and have non-zero values. This is based on the assumption that all features are uncorrelated. However this assumption leads to problems when the transformed models are used. Explicit modelling showed that the diagonal elements of Σ_{kids} and Σ are very similar. For that reason in the rest of the experiments only the means are transformed while the variances are not. It can be argued ([8]) that the variances of original and transformed models are very similar.

Notations and Results Recognition-tests were carried out for the kids test-set using models with the following five different sets of statistics :

1. models trained with the kids training set : μ_{kids}, Σ_{kids}
2. models trained with the adults training set : μ, Σ
3. linear transformations :

$$\begin{aligned} \tilde{\mu}^l_c &= a + \mathbf{B}.\mu_c \\ \tilde{\mu}^l_{\Delta c} &= \mathbf{B}.\mu_{\Delta c} \\ \tilde{\mu}^l_{\Delta^2 c} &= \mathbf{B}.\mu_{\Delta^2 c} \end{aligned}$$

4. non-linear transformation :

$$\tilde{\mu}^n_c = \mathbf{C}.\log\left\{\mathbf{T}.\exp\left(\mathbf{C}^{-1}.\mu_c\right)\right\}$$

Table 2 gives the WER for recognition with the kids test set of TI-digits as a function of the different sets of parameters. An important improvement of 56% is achieved when the transformed models are used. Note that the results with the linear approximation are the same as with the exact calculations based on the algorithm in [8]. However major problems are encountered for the calculation of $\tilde{\mu}_{\Delta^2 c}$ where the lack of information about the non-diagonal elements in Σ seems to affect its estimation. The same was observed when the transformed covariances instead of Σ were used. In this case the results were considerably worse because of the assumption that Σ is diagonal. Furthermore, when the mathematically defined pseudo-inverse of matrix **F** was used instead of the matrix **A** proposed in section 3, a deterioration of WER, 2.3% instead of 1.9%, was observed.

models used	Word Error Rate
kidsmodels ($\mu_{kids}, \Sigma_{kids}$)	1.0 %
adultmodels (μ, Σ)	4.3 %
$\tilde{\mu}^l_c, \tilde{\mu}^l_{\Delta c}, \tilde{\mu}^l_{\Delta^2 c}, \Sigma$	1.9 %
$\tilde{\mu}^n_c, \tilde{\mu}^l_{\Delta c}, \tilde{\mu}^l_{\Delta^2 c}, \Sigma$	1.9 %

Table 2: Recognition results with the kids test set of TI-digits

5. CONCLUSIONS

The feature transformation which is proposed in this paper was shown to be rather efficient in adapting the statistics of a speaker group, for speech recognition purposes. This transformation is successfully linearized when warping is reasonably small. This in turn allowed a better insight into the properties of the transformation, namely with respect to the assumptions related to the diagonal form of the model's covariance matrix. It was shown that models for different classes of speakers have the same variances and only the means need to be transformed. Recognition results showed a clear improvement of more than 50% when the transformed models instead of the original ones were used.

REFERENCES

1. H. Wakita. Estimation of Vocal Tract Shapes from Acoustical Analysis of the Speech Wave: the State of the Art In *IEEE ASSP*, Vol. 27, pp.281, 1979.
2. G. Fant. Acoustic Theory of Speech Production. *The Hague, The Netherlands: Mouton, 1960*
3. G.E.Peterson and H.L.Barney. Control methods Used in a Study of the Vowels In *the Journal of the Acoustical Society of America*, Vol. 24, No. 2 pp.175–184, 1952.
4. E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Int. Conf. Acoust. Speech & Signal Processing*, pages 346–348, 1996.
5. J.G. Wilpon and C.N. Jacobsen. A study of speech recognition for children and the elderly. In *Int. Conf. Acoust. Speech & Signal Processing*, 1996.
6. S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP–28, No. 4 pp.357–366, 1980.
7. G. Fant. Speech Sounds and Features. Cambridge MA: The MIT Press., 1973.
8. T. Claes, I. Dologlou, L. ten Bosch, and D. Van Compernolle. A novel feature transformation for vocal tract length normalisation in automatic speech recognition. Paper submitted to IEEE Transactions on Speech and Audio Processing.