A Novel Feature Transformation for Vocal Tract Length Normalization in Automatic Speech Recognition

Tom Claes, Ioannis Dologlou, Louis ten Bosch, and Dirk Van Compernolle, Member, IEEE

Abstract—This paper proposes a method to transform acoustic models that have been trained with a certain group of speakers for use on different speech in hidden Markov model based (HMM-based) automatic speech recognition. Features are transformed on the basis of assumptions regarding the difference in vocal tract length between the groups of speakers. First, the vocal tract length (VTL) of these groups has been estimated based on the average third formant $F_3$. Second, 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 nonlinear submappings. By locally linearizing it and comparing results in the output, a linear approximation for the exact mapping was obtained which is accurate as long as warping is reasonably small.

The feature vector, which is computed from a speech frame, consists of the mel scale cepstral coefficients (MFCC) along with delta and delta²-cepstra as well as delta and delta² energy. The method has been tested for TI digits data base, containing adult and children speech, consisting of isolated digits and digit strings of different length. The word error rate when trained on adults and tested on children with transformed adult models is decreased by more than a factor of two compared to the nontransformed case.

Index Terms—Speaker adaptation, speaker normalization, vocal tract length normalization.

I. 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 from literature [1]–[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). As a matter of fact, the VTL of females is about 10% shorter compared to the VTL of males. The VTL of children is even shorter (up to 10%) than that of females. According to the linear acoustic theory of speech production, this directly implies that all the formants in male speech undergo a (fixed, VTL-dependent) scaling toward the high end of the spectrum. Consequently, one has to warp the children spectra toward the lower end in order to match it with the adults (male) speech. In this paper, the problem of how to do so is addressed.

Another important issue that is also discussed here is related to the estimation of the VTL from a given speech signal. This is used subsequently to determine the frequency warping factor. It is known that the VTL is related to the position of formants and in particular to the position of the third formant ($F_3$). As opposed to the values of $F_1$ and $F_2$, the value of $F_3$ is less influenced by the vowel under consideration while its detection from the signal itself remains quite reliable. For that reason the estimation of both the VTL and the warping factor are based on the $F_3$ values.

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 delta²-cepstra, and delta-log (energy), delta²-log (energy), i.e., 38 parameters in total.

MFCC parameters [8] are calculated as given in Fig. 1. 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 1 kHz and logarithmically 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. A handy matrix representation is used for all linear steps of the parameter calculation algorithm.

First, a nonlinear transformation relating the MFCC of children 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 (HMM's) takes place. It is shown that the 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 of the TI-digits data base (available through the Linguistic Data Consortium CD no. LDC93S10). Tests are performed using continuous density HMM’s with one model per word (11 states).

II. WARPING TECHNIQUE FOR VOCAL TRACT LENGTH NORMALIZATION

Paper [6] discusses the problems for automatic recognition of child speech. The physiological reasons that change the voice characteristics for males during puberty can be summarized as follows:

- an enlargement of the glottis, which lowers the average pitch;
- a change of the vocal tract length (VTL), which influences the average formant frequencies.

The difference in average VTL for male, female, and children is an important problem for speech recognition with children compared to speech recognition with adults. The VTL is related to the $i$th average formant with (mean) frequency $F_i$ as [1]

$$VTL \approx \frac{(2i-1)c}{4F_i}$$  \hspace{1cm} (1)

where $c$ is the velocity of sound.

According to (1)

$$\frac{VTL_{\text{adults}}}{VTL_{\text{children}}} \approx \frac{F_i_{\text{children}}}{F_i_{\text{adults}}}.$$  \hspace{1cm} (2)

Therefore, normalization of the VTL can be achieved by a frequency warping procedure.

According to [6], MFCC coefficients [8] give better recognition performance with children than LPC’s, because of the mel scale frequency warping. However, the mismatch between male, female and children remains because a VTL dependent warping is necessary to obtain VTL-normalized speech parameters. Most papers indicate that the use of frequency warping procedures improves speaker independent speech recognition performance with males and females [4], [7]. Lee and Rose [7] gives an efficient frequency warping method for the calculation of normalized MFCC’s. This paper combines the concept of frequency warping and the use of MFCC’s [4], [7] and translates them into a transformation of the cepstra.

A. Mathematical Warping Representation

A linear frequency warping of a power-spectrum with $N_{\text{DFT}}$ points can be described with a matrix $W$ whose elements are

$$w(i, j) = \begin{cases} 1, & \text{if } i = \text{round}(w \cdot j) \\ 0, & \text{otherwise} \end{cases}$$

where $w$ is the warping factor (scaling factor) and $1 \leq i, j \leq N_{\text{DFT}}$. If $w > 1$, as in the case of the warping from adults to children, $W$ has the following form:

$$W_{N_{\text{DFT}} \times N_{\text{DFT}}} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}.$$

In return, given the power spectrum $P_1$, $P_2 = WP_1$ is a linearly warped power spectrum by a factor $w$. In other words the frequency-axis is linearly scaled by $w$.

B. The Warping Factor $w$

The main problem with the frequency warping techniques is to determine the warping factor as a function of the VTL. References [1] and [4] propose to use the (mean) third formant ($F_3$). The formula of (2) is more accurate for higher formants. However, formants higher than $F_3$ are difficult to estimate. As a compromise, the warping factor $w$ used here, is defined as the ratio of the average $F_3$ of the speaker class to be recognized to the average $F_3$ of the training class.

The following algorithm is proposed for the calculation of the 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.

The $F_3$ formant-estimator, which is used here, is described below.

1) Perform a peak-detection on the smoothed LPC spectrum.
2) An estimated value for $F_3$ is reliable if:
   a) The frame is a speech frame (speech/silence test)
   b) $F_1 > 250$ Hz ($F_1$: first formant, $\cdots$)
   c) 1700 Hz $< F_3 < 4000$ Hz
   d) $r_1/r_0 > 0.6$ ($r_0$ and $r_1$ are the zero and first order autocorrelation coefficients, respectively).
3) The median is calculated based on frames that satisfy the criteria of Step 2.

As a matter of fact the proposed estimator is not a formant-tracker since only a good guess of $F_3$ is required.

III. MEL CEPSRTUM TRANSFORMATION

In this section, a novel transformation of the mel cepstrum is derived based on the previously presented warping technique. Equation (3) describes the filterbank used to calculate the mel scale spectrum from the DFT power spectrum. The matrix $F$ as defined in (3), shown at the bottom of the next page, is not square and not invertible. The weights $w_{i,n,j}$ are triangular for each output-channel $i$. Above 1 kHz, the channels become wider, i.e., $N_{nc} \gg N_1$, where $N_{nc}$ is the number of channels in the filterbank. The warping method discussed for VTL-normalization should be applied to the linear frequency axis (e.g., power spectrum) and it just represents a scaling of this axis. Note that the exact power spectrum on the linear
frequency axis can not be calculated from the mel scale spectrum $S$, because $F$ is not invertible. A straightforward remedy to this problem is to use the pseudo-inverse of $F$. However, it was observed that this solution does not preserve the formants due to a smoothing effect. Instead an alternative approach is proposed to deduce the spectrum on a linear frequency scale that provides a better approximation after warping and filterbank. This algorithm is described below.

For $i = 1, \ldots, n_c$:
- let $S_i$ be the energy of output-channel $i$ of the filterbank;
- set to $S_i$ the DFT-point which corresponds to the center of channel $i$;
- set the other frequency points of the DFT equal to zero.

This algorithm may be represented by the multiplication of $S$ with matrix $A$ such that $F \cdot A = I$. Note that the product $A \cdot S$ provides the power spectrum where warping applies (simple multiplication with $W$). As a result the complete transformation for the mel-spectrum $S$ can be expressed using the following matrix: $T$:

$$T = F \cdot W \cdot A.$$ 

To deduce the transformation of the mel cepstrum $c$, let $C$ be the matrix representing the DCT transform, the mel-spectrum $S$ corresponding to $c$ is given by $S = \exp(C^{-1} \cdot c)$, in which $\exp(u)$ denotes a components-exponentiation. 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 $\hat{c}$:

$$\hat{c} = C \cdot \log(T \cdot \exp(C^{-1} \cdot c)).$$  \hspace{1cm} (4)

IV. USE OF THE MEL CEPSTRUM TRANSFORMATION FOR HMM ADAPTATION

The transformation of the MFCC can be used to adapt the HMM parameters to the new speaker group characteristics. Suppose that models for a speaker class (e.g., adults) with a set of Gaussians with means $\mu$ and covariance $\Sigma$ are available. The aim is to find the HMM parameters that model speech of another speaker class (e.g., children). In speech recognition applications, both static and dynamic parameters are used, hence $\mu$ and $\Sigma$ may be decomposed as follows:

$$\mu = \begin{bmatrix} \mu_e \\ \mu_{\Delta e} \\ \mu_{\Delta^2 e} \end{bmatrix}$$  \hspace{1cm} (5)

$$\Sigma = \begin{bmatrix} \Sigma_e & \Sigma_{\Delta e} & \Sigma_{\Delta^2 e} \\ \Sigma_{\Delta e} & \Sigma_{\Delta^2 e} & \Sigma_{\Delta^3 e} \\ \Sigma_{\Delta^2 e} & \Sigma_{\Delta^3 e} & \Sigma_{\Delta^4 e} \end{bmatrix}$$  \hspace{1cm} (6)

where $\Delta e$ and $\Delta^2 e$ are defined as:

$$\Delta e(i) = c(i) - c(i-1)$$  \hspace{1cm} (7)

$$\Delta^2 e(i) = \Delta e(i) - \Delta e(i-1)$$  \hspace{1cm} (8)

and the superscripts $(i)$ correspond to the frame-number (time-index).

A. Theoretical Derivation of the Transformed Model Parameters

1) Transformation of the Static Model Parameters: The transformation of a single cepstral vector (static mel cepstra) is calculated using (4). Given the means of the cepstra $\mu_e$ and the covariance matrix $\Sigma_e$, the means $\hat{\mu}_e$ and the covariance matrix $\hat{\Sigma}_e$ of the transformed cepstra is required. This is not a trivial problem because the involved transformation is not linear and therefore

$$E[\hat{c}] = E[C \cdot \log(T \cdot \exp(C^{-1} \cdot c))]$$

$$\neq C \cdot \log(T \cdot \exp(C^{-1} \cdot E[c]))$$

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

$$\hat{\mu}_e \neq C \cdot \log(T \cdot \exp(C^{-1} \cdot \mu_e)).$$

This would only be the case if the transformation was linear.

Later in this section, we will show that because the required warping factors are not large, the exact transformation allows a very good linear approximation that can easily be derived using the Taylor expansion of the exp and log functions.

In the literature [10], an equivalent problem has already been treated and its solution is presented in Fig. 4.

2) Transformation of the Dynamic Model Parameters: The dynamic parameters used for recognition were defined in (7) and (8). Inspired by [10] the statistics of the dynamic parameters of the models may be computed for the transformed delta-cepstra follows:

$$\Delta c(i) = c(i) - c(i-1)$$  \hspace{1cm} (9)

$$\Delta e(i) = c(i) - c(i-1)$$  \hspace{1cm} (10)

$$\Delta^2 e(i) = \Delta e(i) - \Delta e(i-1)$$  \hspace{1cm} (11)
To find the transformed dynamic model parameters, the expectation operator is applied on (11):

\[
E\{\Delta \hat{c}^{(i)}\} = E\{C \cdot \log(T \cdot \exp(C^{-1} \cdot \hat{c}^{(i)}))\} - E\{C \cdot \log(T \cdot \exp(C^{-1} \cdot \hat{c}^{(i)} - \Delta c^{(i)}))\}.
\]  (12)

The computation of the first term in (12) is found in Fig. 4 and equals \(\tilde{\mu}_c\). For the second term first the change of variable \(u = \hat{c} - \Delta c\) is used and thereafter computation proceeds as in Fig. 4. Note that the mean and covariance of \(u\) are

\[
\mu_u = \mu_c - \mu_{\Delta c} \quad \text{and} \quad \Sigma_u = \text{Cov}\{u\} = \Sigma_c + \Sigma_{\Delta c} - 2 \cdot \Sigma_{c, \Delta c}.
\]  (13)

For the delta^2 cepstra the same concept is used.

\[
E\{\Delta^2 c^{(i)}\} = E\{\Delta^2 \hat{c}^{(i)}\} - E\{\Delta^2 \hat{c}^{(i-1)}\} = E\{C \cdot \log(T \cdot \exp(C^{-1} \cdot \hat{c}^{(i)}))\} - E\{C \cdot \log(T \cdot \exp(C^{-1} \cdot \hat{c}^{(i)} - \Delta c^{(i)}))\} + E\{C \cdot \log(T \cdot \exp(C^{-1} \cdot (\hat{c}^{(i-1)} - \Delta c^{(i-1)})))\}.
\]  (15)

The first two terms are the same as in the case of \(\Delta c\), for the third term the change of variable \(v = \hat{c}^{(i)} - 2 \cdot \Delta c^{(i)} + \Delta^2 c^{(i)}\) is proposed and its statistics are given below:

\[
\mu_v = \mu_{\Delta c} - 2 \cdot \mu_{\Delta^2 c} + \mu_{\Delta^2 c} \quad \text{and} \quad \Sigma_v = \text{Cov}\{v\} = \Sigma_{\Delta c} + 4 \cdot \Sigma_{\Delta^2 c} - 4 \cdot \Sigma_{\Delta c, \Delta c} + 2 \cdot \Sigma_{\Delta c, \Delta^2 c}.
\]  (16)

The third term can then be computed using the algorithm in Fig. 4.

**B. Linear Approximation of the Transformation**

In practical applications it is often assumed that only the diagonal elements of \(\Sigma_c\) and \(\Sigma_{\Delta c}\) are nonzero. Since the computation of means and covariances of the transformed parameters based on the algorithm of Fig. 4 makes use of the missing nondiagonal elements their computation may be inaccurate. In order to overcome this problem, a linear approximation to the nonlinear transformation has been derived, and this approximation was shown experimentally to be very accurate. In other words it was discovered that, for reasonably small warping factors, the following holds as a good approximation of the mean of the transformed models.

\[
\hat{\mu}_c \approx C \cdot \log(T \cdot \exp(C^{-1} \cdot \mu_c)).
\]

Furthermore, if \(\hat{c}\) is a vector of MFCC-parameters and \(\hat{c}\) is the transformed vector, it has already been shown that they are related according to (4).

In Appendix A, it is shown by a Taylor expansion how the transformation (4) may be approximated by a linear one of the form

\[
\hat{c} \approx a + B \cdot c
\]  (20)

where

\[
\begin{bmatrix}
\log\left\{ \sum_j t_{1,j} \right\} \\ \log\left\{ \sum_j t_{m,j} \right\}
\end{bmatrix} = C \cdot \begin{bmatrix}
1 \\ 0 \\
0 \\ 1
\end{bmatrix} \cdot T \cdot C^{-1}
\]

and \(t_{i,j}\) are the elements of transformation matrix \(T\).

The accuracy of this approximation will be assessed experimentally in the next chapter through speech recognition experiments.

1) **Transformation of the Static and Dynamic Model Parameters Based on the Linear Approximation:** When the linear approximation (using Taylor expansion) is applied to the mel cepstrum the statistics of the transformed models are given as follows:

\[
E\{\hat{c}\} = E\{C \cdot \log(T \cdot \exp(C^{-1} \cdot c))\} \approx E\{a + B \cdot c\} \approx a + B \cdot E\{c\} \approx a + B \cdot \mu_c
\]

and

\[
\hat{\Sigma}_c \approx B \cdot \Sigma_c \cdot B^t.
\]  (21)

In Appendix B, it is shown how the delta and delta^2 cepstra can be transformed according to

\[
\hat{\Delta} c \approx B \cdot \Delta c
\]

\[
\hat{\Delta}^2 c \approx B \cdot \Delta^2 c
\]

which can be applied directly to the mean values

\[
E\{\hat{\Delta} c\} \approx B \cdot E\{\Delta c\} \quad \hat{\mu}_{\Delta c} \approx B \cdot \mu_{\Delta c} \quad E\{\Delta^2 c\} \approx B \cdot \Delta^2 E\{c\} \quad \hat{\mu}_{\Delta^2 c} \approx B \cdot \mu_{\Delta^2 c}.
\]

The covariances are transformed as

\[
\hat{\Sigma}_{\Delta c} \approx B \cdot \Sigma_{\Delta c} \cdot B^t \quad \hat{\Sigma}_{\Delta^2 c} \approx B \cdot \Sigma_{\Delta^2 c} \cdot B^t.
\]  (22)
V. EXPERIMENTS ON THE TI-DIGITS DATA BASE

In this section the derived transformations are used for speech recognition with children. The goal is to use models that are trained with adult speech for the recognition of children.

A. Data Base Description

All experiments have been carried out using the TI-digits data base.

The specifications of the data are

- sampling frequency 16 kHz (originally recorded at 20 kHz);
- isolated digits and digit strings of different length;
- 22 isolated digits and 55 strings per speaker;
- 55 men, 57 women for training and 56 men, 57 women for testing;
- 25 boys, 26 girls for training and 25 boys, 25 girls for testing (7–14 yrs old).

B. Calculation of the Warping Factor \( w \)

In all transformations discussed in the previous sections the single variable matrix \( W \) is involved. As a matter of fact, this matrix is fully defined as long as the warping factor \( w \) is known. The warping factor related to two speaker classes is defined as the ratio of the mean \( F_3 \) of the classes. Statistics of the \( F_3 \) values were computed both for the training and the testing data of TI-digits. Fig. 2 shows the cumulative distribution functions for each class. The statistics are made from all the utterances in the data base. The average \( F_3 \) and corresponding VTL for each class are given in Table I. The VTL results correspond very well to the results given in [9]. This indicates that the mean third formant \( (F_3) \) is closely related to the VTL. It is therefore clear that speech recognition with children can be improved by normalizing the position of the average \( F_3 \). The warping factor to transform adult models to children models is estimated as follows:

\[
w = \frac{F_{3, \text{children}}}{F_{3, \text{adults}}} = 1.2
\]
with $F_{3, \text{children}}$ and $F_{3, \text{adults}}$ the average $F_3$ for the children and adults class, respectively. In the next section, this factor 1.2 will prove to be sufficiently small to allow the linear approximation of transformation 4 to work.

C. Experimental Comparison of Nonlinear and Linear Transformation

In this section, the linear approximation of (20) is compared with the original transformation of (4). Fig. 3 shows the mean of both a linear and nonlinear transformed cepstral vectors for the case where $w = 1.2$. Note that not only the difference of the means is small but also the variance of the differences is small compared with the absolute value of the mean. This demonstrates experimentally that the approximation

$$C \cdot \log \{T \cdot \exp (C^{-1} \cdot c)\} \approx a + B \cdot c$$

is rather good.

Moreover, it can be concluded that the transformation of (4) is quite linear for small warping factors, and it can be replaced by its linear equivalent.

D. Recognition Experiments

All tests were performed on continuous word recognition using continuous density HMM’s with one model per word (11 states). The training data of the adults (males and females) is used for training the HMM’s. The generated models are represented by their $\mu$ and $\Sigma$. In general only the diagonal elements of $\Sigma$ are known and have nonzero values. This is based on the assumption that all features are uncorrelated. However, this assumption leads to problems when the transformed models are used. To confirm the above assumptions the children training data of TIDigits were used to train models ($\mu_{\text{children}}$ and $\Sigma_{\text{children}}$) and the variances of $\Sigma_{\text{children}}$ were compared with the variances of the adult models $\Sigma$. It was found that the diagonal elements of $\Sigma_{\text{children}}$ and $\Sigma$ are very similar. For that reason in the rest of the experiments only the means are transformed while the variances are not. Besides the experimental confirmation regarding the equality of the variances of adults and children a theoretical proof based on (21) is also proposed. This equation is a similarity transformation which implies that the eigenvalues of $\Sigma_c$ and $\Sigma$ are the same. Since $\Sigma_c$ is assumed to be diagonal, its eigenvalues are equal to the diagonal elements, i.e., the variances. Naturally, since the transformed covariance matrix also corresponds to some uncorrelated cepstral parameters, it has to be diagonal too. Therefore its eigenvalues are also equal to the transformed variances. Since the eigenvalues of $\Sigma_c$ and $\Sigma$ are the same (similarity transformation) the variances of the transformed and the nontransformed parameters are also the same.

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

1) models trained with the children training set $\mu_{\text{children}}, \Sigma_{\text{children}}$;
2) models trained with the adults training set $\mu, \Sigma$;
3) linear transformations

$$\tilde{\mu}_c^l = a + B \cdot \mu_c$$
$$\tilde{\mu}_{\Delta c}^l = B \cdot \mu_{\Delta c}$$
$$\tilde{\mu}_{\Delta^2 c}^l = B \cdot \mu_{\Delta^2 c}$$

4) nonlinear transformation

$$\tilde{\mu}_c^n = C \cdot \log \{T \cdot \exp (C^{-1} \cdot \mu_c)\};$$

5) exact transformation based on the algorithm in Table I:

$$\tilde{\mu} = [\tilde{\mu}_c^l, \tilde{\mu}_{\Delta c}^l, \tilde{\mu}_{\Delta^2 c}^l]'. $$
Computation of the statistics of $\tilde{c}$:

$$\tilde{c} = C \cdot \log \{ T \cdot \exp \{ C^{-1} \cdot c \} \}$$

Definitions: $q \equiv C^{-1} \cdot c$, $r \equiv \exp(q)$, $s \equiv T \cdot r \cdot t \equiv \log(s)$, $\tilde{c} = C \cdot t$

1. $\mu_q = E \{ q \} = C^{-1} \cdot E \{ c \} = C^{-1} \cdot \mu_c$

$$\Sigma_q = C^{-1} \cdot \Sigma_c \cdot (C^{-1})^t$$

2. $\mu_r(i) = E \{ r(i) \} = \exp(\mu_q(i) + \frac{\Sigma_q}{2})$

$$\Sigma_r(i, j) = \mu_r(i) \mu_r(j) (\exp(\Sigma_q(i, j)) - 1) \quad i, j = 1 \ldots n_c$$

3. $\mu_s = E \{ s \} = T \cdot E \{ r \} = T \cdot \mu_r$

$$\Sigma_s = T \cdot \Sigma_r \cdot T^t$$

4. $\mu_t(i) = E \{ t(i) \} = \log(\mu_s(i)) - \frac{1}{2} \log(\frac{\Sigma_s(i, i)}{\mu_s(i) \mu_s(i)}) + 1$

$$\Sigma_t(i, j) = \log(\frac{\Sigma_s(i, j)}{\mu_s(i) \mu_s(j)}) + 1 \quad i, j = 1 \ldots n_c$$

5. $\tilde{\mu}_c = E \{ C \cdot t \} = C \cdot E \{ t \} = C \cdot \mu_t$

$$\tilde{\Sigma}_c = C \cdot \Sigma_t \cdot C^t$$

$E \{ \}$ denotes the expectation operator

The word error rate (WER) is defined as

$$\text{WER} = \frac{N_{\text{sub}} + N_{\text{ins}} + N_{\text{del}}}{N_{\text{tot}}}$$

with $N_{\text{sub}}$, $N_{\text{ins}}$, and $N_{\text{del}}$ the number of substitutions, insertions, and deletions, respectively, and $N_{\text{tot}}$ the total number of words in the test utterances. Table II gives the WER for recognition with the children 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 of Fig. 4. However, major problems are encountered for the calculation of $\tilde{\mu}_c$, where the lack of information about the nondiagonal elements in $\Sigma$ seems to affect its estimation. The same was observed when the transformed covariances instead of $\Sigma$ were used. In this case, the results were considerably worse because of the assumption that $\Sigma$ is diagonal. Furthermore, when the pseudo-inverse of matrix $F$ was used instead of the proposed matrix $A$, a slight deterioration of WER, 2.3% instead of 1.9%, was observed.

VI. CONCLUSIONS

The feature transformation that is proposed in this paper was shown to be rather efficient in adapting the statistics of a speaker group, for speech recognition purposes. Although nonlinear, this transformation is successfully linearized when warping is reasonably small. This, in turn, allowed 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.

APPENDIX A

LINEAR APPROXIMATION OF ADULTS-TO-CHILDREN TRANSFORMATION

The right-hand expression in

$$\tilde{c} = C \cdot \log\{ T \cdot \exp(C^{-1} \cdot c) \}$$

Table II

<table>
<thead>
<tr>
<th>models used</th>
<th>Word Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>children models ($\mu_{\text{children}}, \Sigma_{\text{children}}$)</td>
<td>1.0 %</td>
</tr>
<tr>
<td>adult models ($\mu, \Sigma$)</td>
<td>4.3 %</td>
</tr>
<tr>
<td>$\tilde{\mu}<em>c$, $\tilde{\mu}</em>{\Delta c}$, $\tilde{\mu}_{\Delta^2 c}$, $\Sigma$</td>
<td>1.9 %</td>
</tr>
<tr>
<td>$\tilde{\mu}<em>c$, $\tilde{\mu}</em>{\Delta c}$, $\tilde{\mu}_{\Delta^2 c}$, $\Sigma$</td>
<td>1.9 %</td>
</tr>
<tr>
<td>$\tilde{\mu}<em>c$, $\tilde{\mu}</em>{\Delta c}$, $\tilde{\mu}_{\Delta^2 c}$, $\Sigma$</td>
<td>1.9 %</td>
</tr>
</tbody>
</table>
will be linearly approximated by using the following Taylor-
expansions
\[
\exp(x) \approx 1 + x
\]
\[
\log(1 + x) \approx x
\]
around \( x = 0 \).

Using the linearization of \( \exp(x) \), one obtains
\[
\hat{c} \approx C \cdot \log(\{T \cdot (I + C^{-1} \cdot c)\})
\]
(26) with
\[
I = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}
\]
(as many ones as rows in \( C \)) with some additional algebra on
(26) the following is derived:
\[
\hat{c} \approx C \cdot \log(\{T \cdot (I + C^{-1} \cdot c)\})
\]
\[
\approx C \cdot \log(\left\{ \begin{array}{c}
\sum_j t_{1,j} \\
\vdots \\
\sum_j t_{n_c,j}
\end{array} \right\} + T \cdot C^{-1} \cdot c)
\]
\[
\approx C \cdot \log(\left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\} + T \cdot C^{-1} \cdot c)
\]
\[
\approx C \cdot \log(\left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\}) + C \cdot \log(\left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\})
\]
\[
\approx C \cdot \left( \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right) + C
\]
\[
\cdot \log(I + \left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\} + T \cdot C^{-1} \cdot c)
\]
(30)
\[
(31) \text{ becomes }
\hat{c} \approx C \cdot \log(\left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\}) + C \cdot \log(I + \left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\})
\]
\[
\approx C \cdot \log(I + \left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\}) + C \cdot \log(\left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\})
\]
\[
\approx C \cdot \log(I + \left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\}) + C \cdot \log(\left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\})
\]
\[
\approx C \cdot \log(I + \left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\}) + C \cdot \log(\left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\})
\]
\[
\approx a + B \cdot c
\]
with
\[
\begin{bmatrix}
\log(\{t_{1,j}\}) \\
\vdots \\
\log(\{t_{n_c,j}\})
\end{bmatrix}
\approx C \cdot \log(I + \left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\}) + C \cdot \log(\left\{ \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right\})
\]
\[
B = C \cdot \left( \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right)
\cdot \left( \begin{array}{c}
\sum_j t_{1,j} \\
0 \\
\sum_j t_{n_c,j}
\end{array} \right)
\]
where \( t_{i,j} \) are the elements of transformation matrix \( T \).

APPENDIX B
TRANSFORMATION OF DELTA AND DELTA\(^2\)-CEPSTRA

The derivation of the mean of the transformed (MFCC)
cepstra is straightforward, because the transformation that is
applied to the cepstra is linear and can just be applied to the
means
\[
E(\hat{c}) = E(\{C \cdot \log(\{T \cdot \exp(C^{-1} \cdot c)\})\})
\]
\[
\approx E(\{a + B \cdot c\})
\]
\[
\approx a + B \cdot E(c)
\]
\[
\approx C \cdot \log(\{T \cdot \exp(C^{-1} \cdot E(c))\})
\]
In the case of the delta and delta\(^2\)-cepstra, the transformation
is not the same as for the cepstra because of the offset \( \alpha \)
\[
\Delta c^{(j)} = c^{(j)} - c^{(j-1)}
\]
\[
\approx a + B \cdot (c^{(j)} - c^{(j-1)})
\]
\[
\approx B \cdot \Delta c^{(j)}
\]
\[
\Delta^2 \vec{c}^{(i)} = \Delta \Delta \vec{c}^{(i-1)} \\
\approx \mathbf{B} \cdot \Delta \vec{c}^{(i)} - \mathbf{B} \cdot \Delta \vec{c}^{(i-1)} \\
\approx \mathbf{B} \cdot (\Delta \vec{c}^{(i)} - \Delta \vec{c}^{(i-1)}) \\
\approx \mathbf{B} \cdot \Delta^2 \vec{c}^{(i)}.
\]

Note that only the scaling matrix \( \mathbf{B} \) is important because the translation \( \vec{a} \) cancels out. The linear transformation can then be applied to the mean-values:

\[
E\{\Delta^2 \vec{c}^{(i)}\} \approx \mathbf{B} \cdot E\{\vec{c}^{(i)}\} \\
E\{\Delta^2 \vec{c}^{(i)}\} \approx \mathbf{B} \cdot E\{\Delta^2 \vec{c}^{(i)}\}.
\]

REFERENCES


Tom Claes was born in Belgium in 1969. He received the degree in electrical engineering from the Katholieke Universiteit Leuven (KUL), Leuven, Belgium, in 1992.

From 1992 until 1996, he worked as Research Assistant for the Speech Group, Electrical Engineering Department, KUL. His main research activities were in signal processing for robust speech recognition in adverse conditions. Since 1997, he has been doing research for automatic speech recognition at Lernout & Hauspie Speech Products, Wemmel, Belgium.

Iooannis Dologlou received the M.S. degree in electrical engineering from the National Technical University of Athens (NTUA), Greece, in 1981, and the Ph.D. degree from the Institut National Polytechnique of Grenoble, France, in 1984.

From 1984 until 1986, he was a Post-doctoral Fellow at the University of California, Santa Barbara, and from 1986 until 1988, he was a Research Assistant at Imperial College, London, U.K. From 1989 until 1991, he was a Research Fellow at NTUA and Consultant for the European Project in automatic translation EUROTRA-EL. During 1992, he was a Visiting Professor at the University of Paris-Dauphine and Paris-Nord, and from 1993 until 1994, he was a Research Director at the French CNRS. In 1995 and 1996, he was a Research Scientist at the Technical University of Delft, The Netherlands, and the Katholieke University of Leuven (KUL), Belgium, respectively. Since October 1996, he has been the Head of the Speech Group, Electrical Engineering Department, KUL. His research interests are in speech processing and SVD-based spectral estimation.

Louis ten Bosch (1957) received the Master’s degree in mathematics from Rijks Universiteit Leiden, The Netherlands, and the Ph.D. degree from the University of Amsterdam, The Netherlands. His dissertation was on phonetic models underlying phonological tendencies in phoneme inventories across languages.

From 1990 to 1992, he worked on automatic rule extraction for text-to-speech methods within the framework of the Project for Advancement of Computer Science and Speech at the Catholic University of Nijmegen and the Institute of Perception Research, Eindhoven, The Netherlands. He also performed research on automatic classification of pitch movements and accent locations. His current research interest is optimization of feature extraction for automatic speech recognition and modeling of human classification strategies.

Dirk Van Compernolle (S’82–M’85) was born in Kortrijk, Belgium, in 1956. He received the electrical engineering degree from the Katholieke Universiteit Leuven (KUL), Heverlee, Belgium, in 1979, and the M.Sc. and Ph.D. degrees from Stanford University, Stanford, CA, in 1982 and 1985, respectively. His doctoral research was on speech signal processing for cochlear implants.

From 1985 until 1987, he was with the IBM T. J. Watson Research Center, Yorktown Heights, NY, where he performed research on robust speech recognition. In 1987, he joined the staff of the Electrical Engineering Department, KUL, where he held positions of Senior Research Scientist of the National Science Foundation of Belgium, Assistant Professor, and part-time Professor since 1994. In June 1994, he joined Lernout & Hauspie (L&H) Speech Products, Wemmel, Belgium, as a Vice President and the Head of the Speech Recognition Division. In 1997, he became Head of the Basic Research Division, L&H. His current research interests include the domains of robust speech recognition, speech enhancement, microphone arrays, auditory modeling, and signal processing for hearing aids. He is a member of the editorial board of the journal Speech Communication.