ELSEVIER

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Speech Communication 42 (2004) 75–91

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Speaker adaptation with all-pass transforms

John McDonough [*], Thomas Schaaf, Alex Waibel

Interactive Systems Laboratories, Institut für Logik, Komplexität, und Deduktionssysteme, Universität Karlsruhe,
Am Fasanengarten 5, 76128 Karlsruhe, Germany

Abstract

Modern speech recognition systems are based on the hidden Markov model (HMM) and employ cepstral features to represent input speech. In *speaker normalization*, the cepstral features of speech from a given speaker are transformed to match the speaker independent HMM. In *speaker adaptation*, the means of the HMM are transformed to match the input speech. Vocal tract length normalization (VTLN) is a popular normalization scheme wherein the frequency axis of the short-time spectrum is rescaled prior to the extraction of cepstral features. In this work, we develop novel speaker adaptation schemes by exploiting the fact that frequency domain transformations similar to that inherent in VTLN can be accomplished entirely in the cepstral domain through the use of conformal maps. We describe two classes of such maps: *rational all-pass transforms* (RAPTs) which are well-known in the signal processing literature, and *sine-log all-pass transforms* (SLAPTs) which are novel in this work. For both classes of maps, we develop the relations necessary to perform maximum likelihood estimation of the relevant transform parameters using enrollment data from a new speaker. We also propose the means by which an HMM may be trained specifically for use with this type of adaptation. Finally, in a set of recognition experiments conducted on conversational speech material from the *Switchboard Corpus* as well as the *English Spontaneous Scheduling Task*, we demonstrate the capacity of APT-based speaker adaptation to achieve word error rate reductions superior to those obtained with other popular adaptation techniques, and moreover, reductions that are *additive* with those provided by VTLN.

© 2003 Elsevier B.V. All rights reserved.

Keywords: Speaker adaptation; Speech recognition

1. Introduction

Associated with each state in a continuous density hidden Markov model (HMM) is a probability density function (pdf). In the absence of any normalization or adaptation, the pdf for a single HMM state s can be expressed as a mixture of Gaussian components:

$$P(x;\Lambda) = \sum_{k=1}^{K} q_k P(x;\Lambda_k) \tag{1}$$

where x is an observation vector, $\{q_k\}$ is the set of a priori probabilities for each of the mixture components, and $P(x;\Lambda_k)$ is the kth Gaussian density function. The latter can be expressed as

$$P(x;\Lambda_k) = \frac{1}{\sqrt{|2\pi D_k|}} \exp\left[-\frac{1}{2}(x-\mu_k)^{\mathrm{T}} D_k^{-1}(x-\mu_k)\right] \tag{2}$$

where μ_k and D_k are the mean and covariance respectively, which together comprise Λ_k.

[*] Corresponding author. Tel.: +49-721-608-4732; fax: +49-721-607-721.
E-mail address: jmcd@ira.uka.de (J. McDonough).
URL: http://isl.ira.uka.de/~jmcd.

0167-6393/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2003.09.005

In *speaker adaptation* we attempt to transform the means of a HMM so as to match the speech from a particular speaker. Most transform-based speaker adaptation techniques employ linear transformations of the type

$$\hat{\mu}_k = A^{(s)}\mu_k + b^{(s)} \tag{3}$$

to obtain the transformed means $\{\hat{\mu}_k\}$ from the initial means $\{\mu_k\}$ of the speaker-independent (SI) model, where $A^{(s)}$ and $b^{(s)}$ are respectively the speaker-dependent transformation matrix and additive bias. Modifying (2), the likelihood assigned an observation x by the kth Gaussian component is

$$P(x; A^{(s)}, b^{(s)}, \Lambda_k) = \frac{1}{\sqrt{|2\pi D_k|}} \exp\left[-\frac{1}{2}(x - \hat{\mu}_k)^{\mathrm{T}} \times D_k^{-1}(x - \hat{\mu}_k)\right]$$

To be effective, the transform parameters $(A^{(s)}, b^{(s)})$ must satisfy two conflicting requirements, these being

(1) the necessity of using a powerful transformation in order to capture the fine differences among speakers;
(2) the need of a transformation specified with few parameters to ensure they can be reliably estimated.

Undoubtedly the most popular formulation of such a speaker adaptation scheme is maximum likelihood linear regression (MLLR) (Leggetter and Woodland, 1995).

In all speaker adaptation schemes, a set of speaker-dependent transformation parameters must be estimated from some amount of *enrollment data*, typically using a maximum likelihood criterion. Parameter estimation is typically performed via the *expectation-maximization* (EM) algorithm, which requires a set of phone-level transcriptions of the utterances on which the estimation is based. In those speaker adaptation scenarios of most immediate interest, no reference transcriptions are provided for the enrollment data, which is to say that the parameter estimation is *unsupervised*. Hence, it is necessary to use an unadapted speech recognition system to form an initial hypothesis for the utterances of a given speaker. This initial hypothesis will thus contain errors, which renders the second criterion above all the more pressing, as such errors make it difficult to reliably estimate large numbers of parameters.

All current state of the art speech recognition systems make use of observation vectors $\{x_i\}$ or *features* composed of *cepstral sequences* (Oppenheim and Schafer, 1989, Section 12.1) and their first and second order differences to represent any input speech. These sequences are the coefficients in a series expansion of an analytic function, and thus contain a great deal of structure—structure that might be exploited when performing speaker adaptation. The speaker adaptation paradigms mentioned above are predicated on linear transformations estimated using a maximum likelihood (ML) criterion; through the artifice of the auxiliary function—the vital center of the EM algorithm—this ML criterion can be reduced to a weighted least squared-error metric. Hence, these techniques are essentially equivalent to linear regression. While effective in the present application, linear regression is a general purpose technique and completely disregards the unique structure of cepstral features which are to be transformed. It is the objective of the present work to turn this structure to good advantage in formulating more effective speaker adaptation paradigms.

Speaker normalization is closely related to speaker adaptation, inasmuch as it attempts to transform the features of a given speaker's speech to match a speaker independent (SI) model. In prior work (McDonough et al., 1998), we explored the use of the bilinear transform (BLT), and a generalization thereof dubbed the all-pass transform (APT), as a means of formulating practical speaker normalization schemes. Two factors were critical in motivating these earlier investigations: Firstly, the BLT approximates to a reasonable degree the frequency domain transformations most often used in vocal tract length normalization (VTLN), which is arguably the most popular and effective speaker normalization technique in use today (Andreou et al., 1994; Pye and Woodland, 1997). Secondly, both the BLT and APT can be represented as linear transformations in the cepstral domain (Acero, 1990). This latter property provides for a straightforward speaker normalization scheme—it is in fact possible to apply speaker normalization *on-the-fly* during training or recognition starting from unnormalized cepstra (McDonough et al., 1998). In addition, the linearity of the underlying transformation lends itself to robust estimation of the requisite speaker dependent transformation parameters (McDonough et al., 1998). Indeed, the advantages afforded by this linearity have been more recently recognized by other authors (Pitz et al., 2001; Ding et al., 2002).

1.1. Review of prior work

Masry et al. (1968) considered the possibility of representing a continuous-time signal as a discrete-time sequence. Their approach to this problem was posed in terms of defining a basis of orthonormal functions that is complete for signals with particular smoothness properties. Oppenheim and Johnson (1972) took (Masry et al., 1968) as their starting point in deriving a class of transformations that preserve convolution. They found that one of the principal requirements for such a class is that it have the form of the composition of two functions. Oppenheim and Johnson (1972) also developed the mathematical basis for using the BLT to transform discrete-time sequences, and showed this transformation could be accomplished via a cascade of first order difference equations.

Zue (1971) used the technique of (Oppenheim and Johnson, 1972) to restore the speech of divers breathing helium-rich gas mixtures. Shikano (1986, Section 7) noticed the similarity of the BLT to the mel-scale and used it to apply a speaker-independent warp to the short-time spectrum of speech prior to recognition. Acero (1990, Section 7), first proposed using a speaker-dependent BLT to correct for inter-speaker differences in formant frequency locations; in this work, the optimal BLT parameter for each speaker was estimated by minimizing a vector quantization distortion measure.

After lying dormant for several years, the use of VTLN to enhance the performance of large vocabulary conversational speech recognition (LVCSR) systems was re-introduced by Andreou et al. (1994). Their technique had no recourse to the BLT. Instead, a speech waveform was sampled at various rates to induce a linear scaling on the frequency axis of the short-time Fourier transform; the final sampling rate for a particular speaker was chosen to minimize the number of errors made by an HMM-based LVCSR system. The publication of (Andreou et al., 1994) sparked a flurry of activity: Eide and Gish (1996) proposed a nonlinear warping of the short-time frequency axis implemented in the spectral domain; the choice of warp factor was based on explicit estimates of speaker-dependent formant frequencies. Wegmann et al. (1996) and Lee and Rose (1996) independently proposed the use of a Gaussian mixture model to obtain ML estimates of the optimal warping parameters. Pye and Woodland (1997) investigated the use of VTLN together with MLLR adaptation; their findings indicated that the reductions in word error rate achieved by VTLN and MLLR when used in isolation were largely additive when these techniques were combined.

Digalakis et al. (1995) introduced transformed-based adaptation of Gaussian mixtures. In this scheme, the kth Gaussian mean was transformed as in (3) where $A^{(s)}$ was taken as diagonal, and the kth covariance matrix Σ_k was transformed as

$$\widehat{\Sigma}_k = A^{(s)}\Sigma_k A^{(s)}t$$

Hence, the transformation applied to the covariance matrix was completely determined by that applied to the mean; for this reason, the approach of (Digalakis et al., 1995) came to be known as a *constrained adaptation* of Gaussian mixtures.

Leggetter and Woodland (1995) proposed the highly successful MLLR adaptation. Their technique was similar to that of (Digalakis et al., 1995) in that the means of a speaker-independent HMM were transformed as in (3), but differed in that $A^{(s)}$ was taken as a full, instead of diagonal, matrix. In this initial work, only the Gaussian mean was transformed; a covariance transform was subsequently added by Gales and Woodland (1996). In the latter work, the transform applied to the covariance matrix was not explicitly tied to that applied to the mean; hence, this was the first instance of what came to be known as an *unconstrained adaptation*.

The adaptation techniques mentioned above all transform a conventionally-trained speaker-independent model. Anastasakos et al. (1996) first considered the possibility of training a speaker-independent HMM specifically for use with speaker adaptation. In their technique, transform parameters are first estimated for all speakers in a training set. Then the Gaussian means and variances of a speaker-independent HMM are iteratively re-estimated using the transform parameters of the training set speakers along with the usual forward-backward statistics.

An excellent review of the aforementioned transformation-based approaches to speaker adaptation, along with the requirements of each in terms of computation and memory, is given by Gales (1998). Another valuable reference is Sankar and Lee (1996), who formulate a unified basis for ML speaker normalization and adaptation.

Recently there has been a growing interest in performing speaker adaptation with very limited amounts of enrollment data; e.g., 30 s or less. The results of some preliminary investigations in this area have been reported by Digalakis et al. (1996), by Kannan and Khudanpur (1999), and by Bocchieri et al. (1999). A distinctly different approach to the problem of rapid adaptation is formulated by Gunawardana and Byrne (2000); it involves the use of a *discounted likelihood* criterion to achieve robust parameter estimation. Another popular and effective approach to very rapid adaptation, dubbed *eigenvoices*, was developed by Kuhn et al. (2000).

1.2. Organization of this work

Let us outline the balance of this work. The characteristics of the general, rational all-pass transform (RAPT) are presented in Section 2, as is a method by which these functions can be used to transform general discrete-time sequences. This section is based on the work by McDonough (2000), but is not as mathematically rigorous as that earlier publication; for reasons of brevity, there are none of the analyticity arguments to which a great deal of painstaking development was devoted in (McDonough, 2000). Section 2.4 introduces the sine-log all-pass transform (SLAPT), and discusses its computational advantages over the RAPT discussed earlier.

Section 3 discusses the maximum likelihood estimation of APT parameters using a set of enrollment data collected from a new speaker. In this development, the likelihood of the enrollment data is maximized via the EM algorithm. The "engineering" details of applying APTs to speaker compensation are also briefly discussed in Section 3. Of particular interest here is the use of HMMs with a single Gaussian component per state cluster to estimate the parameters of an APT.

Section 4 documents the results of several experiments establishing the capacity of the techniques proposed in this work to improve speech recognition performance. Of particular interest in this section is the empirical demonstration that SLAPT adaptation provides word error rate reductions superior to those given by MLLR, and that these reductions are additive with those achieved by VTLN.

Finally, Section 5 summarizes what we have learned about speaker adaptation with all-pass transforms, and suggests ways in which this approach might be extended in future.

2. Theoretical development

Here we set forth the characteristics of a class of mappings which are designated *all-pass transforms* for reasons which will emerge presently. We also describe how these mappings can be used in transforming cepstral sequences.

2.1. Sequence transformation

Consider an arbitrary double-sided, real-valued time sequence $c[n]$ and its z-transform $C(z)$, which are related by the equations

$$C(z) = \sum_{n=-\infty}^{\infty} c[n] z^n \tag{4}$$

$$c[n] = \frac{1}{2\pi j} \oint C(z) z^{-(n+1)} \, \mathrm{d}z \tag{5}$$

where the contour of integration in (5) is assumed to be the unit circle. This non-standard definition of a z-transform pair is used here to facilitate the development that follows.

For some mapping Q, assume we wish to form the composition $\hat{C}(z) = C(Q(z))$. If Q satisfies suitable analyticity conditions, then $\hat{C} = C \circ Q$ also admits a Laurent series representation

$$\hat{C}(z) = \sum_{n=-\infty}^{\infty} \hat{c}[n]z^n$$

McDonough (2000), showed that the series coefficients $\hat{c}[n]$ appearing above can be calculated from

$$\hat{c}[n] = \sum_{m=-\infty}^{\infty} c[m]q^{(m)}[n] \tag{6}$$

where

$$q^{(m)}[n] = \frac{1}{2\pi j}\oint Q^m(z)z^{-(n+1)}\,\mathrm{d}z \tag{7}$$

Furthermore, the several sequences $\{q^{(m)}[n]\}$ satisfy

$$q^{(m)}[n] = \sum_{k=-\infty}^{\infty} q^{(m-1)}[n]q^{(1)}[n-k] \tag{8}$$

and $q^{(0)}[n]$ is equivalent to the unit sample sequence:

$$q^{(0)}[n] = \begin{cases} 1, & \text{for } n = 0 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

From (8) and (9) it is clear the sequences $\{q^{(m)}[n]\}$ for all $m = 2, 3, \ldots$ can be readily calculated once $q^{(1)}[n] = q[n]$ is known. In the following sections, we show how $q[n]$ can be obtained for both rational and sine-log all-pass transforms.

2.2. Rational all-pass transforms

Here we propose a class of functions $\{Q\}$ that can be used to adapt sequences of cepstral coefficients. Much of this development follows the classic work of Oppenheim and Johnson (1972). Consider the well-known *bilinear transform* (BLT), which for the present purpose will be expressed as

$$Q(z) = \frac{z-\alpha}{1-\alpha z} \tag{10}$$

for some real α. The effect of the BLT can be equated to a non-linear warping of the frequency axis (Oppenheim and Johnson, 1972). Indeed, defining the transformed angular frequency $\omega' = \arg Q(e^{j\omega})$, and working directly from (10), it is straightforward to show

$$\omega' = \tan^{-1}\frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha}$$

The resulting plot of ω' versus ω is shown in Fig. 1, from which it is apparent that the frequency axis can be warped up or down through suitable settings of α, much as in traditional vocal tract length normalization (VTLN).

From the development in the preceding section, it is clear that $\hat{c}[n]$ can be readily calculated as soon as the coefficients q in the series expansion of Q are known. To calculate the latter we begin with the well-known *geometric series*,

$$\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n$$

which holds for all $|z| < 1$. Using this series, it is possible to rewrite $Q(z)$ as

$$Q(z) = (z-\alpha)\sum_{n=0}^{\infty}\alpha^n z^n$$

for all $|z| < \alpha^{-1}$. From this equation, the individual coefficients of the series expansion of Q can be determined by inspection:

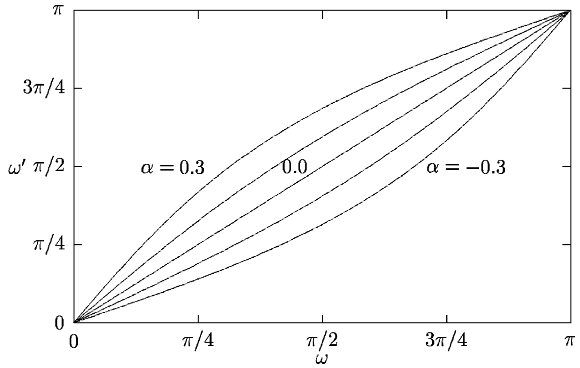

Fig. 1. Effect of the bilinear transform on the mapped frequency $\omega' = \arg Q(e^{j\omega})$ for various values of α. Setting $\alpha = 0$ corresponds to the identity transformation.

$$q[0] = -\alpha \tag{11}$$

$$q[n] = \alpha^{n-1}(1-\alpha^2) \quad \text{for all } n > 0 \tag{12}$$

It is possible to formulate a more general class of mappings that share many of the desirable characteristics of the bilinear transform. The mappings are dubbed *rational all-pass transforms* (RAPTs), and have the functional form:

$$\begin{aligned} Q(z) &= \underbrace{\frac{z-\alpha}{1-\alpha z}}_{} \times \underbrace{\frac{z-\beta}{1-\overline{\beta}z}\,\frac{z-\overline{\beta}}{1-\beta z}}_{} \times \underbrace{\frac{1-\overline{\gamma}z}{z-\gamma}\,\frac{1-\gamma z}{z-\overline{\gamma}}}_{} \\ &= A(z;\alpha) \times B(z;\beta) \times G(z;\gamma) \end{aligned} \tag{13}$$

where α, β, $\gamma \in C$ satisfy $|\alpha|$, $|\beta|$, $|\gamma| < 1$. From (10) and (13) it is apparent that the latter mapping subsumes the former, and the two are equivalent whenever $\beta = \gamma$. In the sequel, the dependence of $A(z;\alpha)$, $B(z;\beta)$, and $G(z;\gamma)$ on α, β, and γ shall be suppressed whenever it is possible to do so without ambiguity.

The general RAPT in (13) has several salient characteristics, two of which we now enumerate:

(1) Q is an all-pass function such that

$$|Q(e^{j\omega})| = 1 \tag{14}$$

for all $\omega \in R$;

(2) The inverse of Q is available from

$$Q(z^{-1}) = \frac{1}{Q(z)} \tag{15}$$

Discrete-time systems having transfer functions that can be represented as a product of terms of the type seen in (13) are frequently used for phase compensation of digital filters (Oppenheim and Schafer, 1989, Section 5.5). As implied by (14), cascading a phase compensator of this type with an arbitrary linear time-invariant filter does not alter the spectral *magnitude* of the latter. For this reason, such a system is described as *all-pass*; that is, it passes all frequencies without attenuation. In the sequel, we will use the term *all-pass transform* to refer to any conformal map satisfying conditions (14) and (15).

Two observations can be made based on the foregoing: The placement of poles and zeros in (13) is dictated by the *argument principle* (Churchill and Brown, 1990). In particular, we require that the number of zeros within the unit circle exceed the number of poles by exactly *one*. Moreover, as a consequence of condition (14), the effect of any APT can be equated to a non-linear warping of the frequency axis, just as was previously done with the BLT. Details are provided in (McDonough, 2000, Section 3.2).

Suppose that Q is an RAPT as in (13) and that $|\alpha|$, $|\beta|$, and $|\gamma| < 1$. Then Q admits a Laurent series representation

$$Q(z) = \sum_{n=-\infty}^{\infty} q[n]z^n \tag{16}$$

whose coefficients are given by

$$q = a * b * g \tag{17}$$

where $a \leftrightarrow A$, $b \leftrightarrow B$, and $g \leftrightarrow G$. The components of a were given in (11) and (12). Comparable expressions for the components of b and g are derived in (McDonough, 2000).

As we are transforming a cepstral sequence $c[n]$ which is inherently double-sided, it is necessary to calculate $q^{(m)}[n]$ for both positive and *negative* integers m. We can, however, exploit the special structure of the APT in order to relate the components of $q^{(m)}$ to those of $q^{(-m)}$ for $m \geqslant 1$. Note that

$$Q^{-m}(z) = \left[\frac{1}{Q(z)}\right]^m = [Q(z^{-1})]^m$$

where the final equality follows from (15). Hence,

$$Q^{-m}(z) = Q^m(z^{-1})$$

which implies

$$q^{(-m)}[n] = q^{(m)}[-n] \tag{18}$$

The import of Eq. (18) is that only the set of sequences $\{q^{(m)}[n]\}$ for all $m \geqslant 0$ need be calculated directly, and (8) provides the means to accomplish this once $q^{(1)}[n] = q[n]$ is known; the latter is available from (17).

2.3. Cepstral sequence transformation

As nearly all modern speech recognizers use cepstral sequences as input features, we must specialize the development above for the unique characteristics of cepstral coefficients. Hence, define $\hat{X}(z) = \log \hat{H}(z)$ and $X(z) = \log H(z)$ so that $\hat{X} = X \circ Q$. If $c[n]$ is the real cepstrum corresponding to some windowed segment of speech then $c[n]$ must be even. Define $x[n]$ as the *minimum phase* (Oppenheim and Schafer, 1989, Chapter 12) equivalent of $c[n]$, such that

$$x[n] = \begin{cases} 0; & \text{for } n < 0 \\ c[0]; & \text{for } n = 0 \\ 2c[n]; & \text{for } n > 0 \end{cases}$$

and

$$c[n] = \begin{cases} \frac{1}{2}x[-n], & \text{for } n < 0 \\ x[0], & \text{for } n = 0 \\ \frac{1}{2}x[n], & \text{for } n > 0 \end{cases} \tag{19}$$

Exploiting the fact that $c[n]$ is even, rewrite (6) as

$$\begin{aligned} \hat{c}[n] &= q^{(0)}[n]c[0] + \sum_{m=1}^{\infty} \left(q^{(m)}[n] + q^{(-m)}[n]\right)c[m] \\ &= q^{(0)}[n]c[0] + \sum_{m=1}^{\infty} \left(q^{(m)}[n] + q^{(m)}[-n]\right)c[m] \end{aligned}$$

Substituting (19) into the last expression then gives

$$\hat{c}[n] = q^{(0)}[n]x[0] + \frac{1}{2}\sum_{m=1}^{\infty} \left(q^{(m)}[n] + q^{(-m)}[n]\right)x[m] \tag{20}$$

Now define $\hat{x}[n]$ as the causal portion of $\hat{c}[n]$, so that

$$\hat{x}[n] = \begin{cases} 0; & \text{for } n < 0 \\ \hat{c}[0]; & \text{for } n = 0 \\ 2\hat{c}[n]; & \text{for } n > 0 \end{cases} \tag{21}$$

Substituting (20) into (21) provides

$$\hat{x}[0] = \sum_{m=0}^{\infty} q^{(m)}[0]x[m] \tag{22}$$

and

$$\hat{x}[n] = \sum_{m=1}^{\infty} \left(q^{(m)}[n] + q^{(m)}[-n]\right)x[m] \tag{23}$$

for all $n > 0$. These relations can be stated more succinctly by defining the *transformation matrix* $A = \{a_{nm}\}$ where

$$a_{nm} = \begin{cases} q^{(m)}[0], & \text{for } n = 0, \ m \geqslant 0 \\ 0, & \text{for } n > 0, \ m = 0 \\ (q^{(m)}[n] + q^{(m)}[-n]), & \text{for } n, \ m > 0 \end{cases} \tag{24}$$

Hence, it is possible to obtain $\hat{x}[n]$ from

$$\hat{x}[n] = \sum_{m=0}^{\infty} a_{nm}x[m] \tag{25}$$

From (25) it is clear that the composition $\hat{X} = X \circ Q$ reduces to a linear transformation in cepstral space. It is worth noting that this is a consequence of the fact that Q

(1) is analytic on an annular region that includes the unit circle, and
(2) preserves the unit circle.

The claims of Pitz et al. (2001) notwithstanding, composition with any function that fails to satisfy either of these requirements will *not* reduce to a linear transformation in cepstral space. This can be readily seen from the following argument. A cepstral sequence is comprised of the coefficients of a Laurent series defined on the unit circle, and hence defines a (unique) function that is analytic on an annular region that includes the unit circle. Moreover, *only* functions that are analytic on a given annular region possess Laurent series representations on that region. Pitz posits no condition of analyticity on his warping functions. Moreover, he considers warping functions that are *piecewise* linear. An analytic function possesses continuous derivatives of *all* orders, which implies the piecewise warping functions that Pitz considers are clearly not analytic as their first derivative is undefined at the point of transition from one linear segment to another. Nor is the composition of an analytic function and a piecewise continuous function analytic, a fact easily verified with the chain rule. Thus, Pitz compositions are not analytic, and therefore possess no valid series representations; i.e., they do not yield valid cepstra.

Fig. 2 shows the original and transformed spectra for a windowed segment of male speech sampled at 8 kHz; both spectra were generated from the first 15 components of the original cepstral sequence. The operations employed in calculating the transformed cepstra $\hat{x}[n]$ were those set forth above. The conformal map used in this case was a bilinear transform with $\alpha = 0.10$. As implied by (25), some of the information contained in $x[n]$ for $n = 0, 1, \ldots, N-1$ is "encoded" in $\hat{x}[n]$ for all $n \geqslant N$; thus, 25 rather than 15 transformed cepstral coefficients were retained in generating the composite spectrum plotted in the figure. It is clear from a comparison of the respective spectra that all formants have been shifted downward by the transformation and that the extent of the shift is frequency dependent. Qualitatively, this is just what we should expect based on the curves plotted in Fig. 1.

Shown in Fig. 3 are the original and transformed spectra for the same segment of male speech previously plotted in Fig. 2. As in the prior case, these plots were generated from the first 15 components of the original cepstral sequence, but 25 components were retained in the transformed sequence. The latter was obtained in the manner suggested by the development above. The conformal map used in this instance was an APT with the general form in (13). From the figure it is apparent that whereas the higher formants have been shifted *down*, the lower formants have been shifted *up*. This stands in sharp contrast to the effect produced by the BLT, for which the shift depends on frequency but is always in the same direction, and serves to illustrate the greater power and generality of the APT.

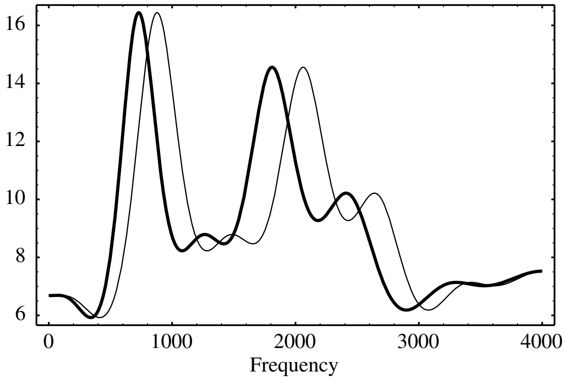

Fig. 2. Original (thin line) and transformed (thick line) short-term spectra for a male test speaker regenerated from cepstral coefficients 0–14. The transformed spectrum was produced with the BLT by setting $\alpha = 0.10$.

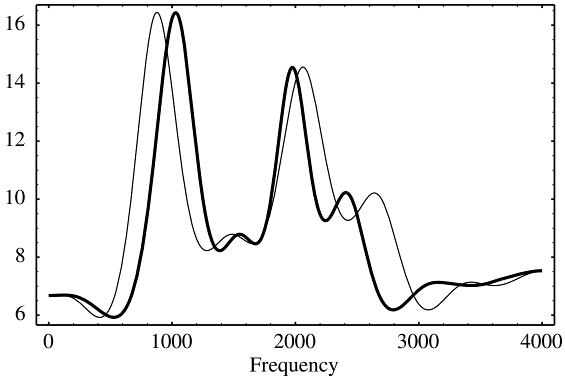

Fig. 3. Original (thin line) and all-pass (thick line) transformation of the short-term spectrum for a male test speaker generated from cepstral coefficients 0–14. The composite spectrum was obtained with an all-pass transform as given in (13).

2.4. Sine-log all-pass transforms

In the final portion of this section, we consider a different type of all-pass transform that shares many of the characteristics of the RAPT. Its chief advantage over the RAPT is its simplicity of form and amenability to numerical computation. Regrettably, this simplicity is not immediately apparent from the abbreviated presentation given here. The interested reader is referred to McDonough (2000) for further details.

Let us begin by defining the *sine-log all-pass transform* (SLAPT) as

$$Q(z) = z \exp F(z) \tag{26}$$

where

$$F(z) = \sum_{k=1}^{K} \alpha_k F_k(z) \quad \text{for } \alpha_1, \ldots, \alpha_K \in R, \tag{27}$$

$$F_k(z) = j\pi \sin\left(\frac{k}{j} \log z\right) \tag{28}$$

and K is the number of free parameters in the transform. The designation "sine-log" is due to the functional form of F_k. It is worth noting that $F_k(z)$ is single-valued even though $\log z$ is multiple-valued. Moreover, applying the well-known relation

$$\sin z = \frac{1}{2j}(\mathrm{e}^{jz} - \mathrm{e}^{-jz})$$

to (28) provides

$$F_k(z) = \frac{\pi}{2}(z^k - z^{-k}) \tag{29}$$

which is a more tractable form for computation. It can be readily verified that Q as defined (26) satisfies (14) and (15) just like the rational APTs considered earlier. Moreover, as z traverses the unit circle, $Q(z)$ also winds exactly *once* about the origin, which is necessary to ensure that spectral content is not doubled or tripled (McDonough, 2000, Section 3.5).

In order to calculate the coefficients of a transformed cepstral sequence in the manner described above, it is first necessary to calculate the coefficients q in the Laurent series expansion of Q; this can be done as follows: For F as in (29) set

$$G(z) = \exp F(z) \tag{30}$$

and let g denote the coefficients of the Laurent series expansion of G valid in an annular region including the unit circle. Then,

$$g[n] = \frac{1}{2\pi j} \oint G(z) z^{-(n+1)}\,\mathrm{d}z \tag{31}$$

The natural exponential admits the series expansion

$$\mathrm{e}^z = \sum_{m=0}^{\infty} \frac{z^m}{m!}$$

so that

$$G(z) = \sum_{m=0}^{\infty} \frac{F^m(z)}{m!}$$

Substituting the latter into (31) provides

$$\begin{aligned} g[n] &= \frac{1}{2\pi j} \oint \sum_{m=0}^{\infty} \frac{F^m(z)}{m!} z^{-(n+1)}\,\mathrm{d}z \\ &= \sum_{m=0}^{\infty} \frac{1}{m!} \frac{1}{2\pi j} \oint F^m(z) z^{-(n+1)}\,\mathrm{d}z \end{aligned} \tag{32}$$

The sequence f of coefficients in the series expansion of F are available by inspection from (27) and (29). Letting $f^{(m)}$ denote the coefficients in the series expansion of F^m, we have

$$f^{(m)}[n] = \frac{1}{2\pi j} \oint F^m(z) z^{-(n+1)}\,\mathrm{d}z$$

and upon substituting this into (32) we find

$$g[n] = \sum_{m=0}^{\infty} \frac{1}{m!} f^{(m)}[n]$$

From the Cauchy product it follows

$$f^{(m)} = f^{(1)} * f^{(m-1)}$$

for $m = 1, 2, 3, \ldots$ Eq. (30) implies that $Q(z) = zG(z)$, so the desired coefficients are given by

$$q[n] = g[n-1]$$

for all $n = 0, \pm 1, \pm 2, \ldots$

3. Practical speaker compensation

Having presented all the theory necessary to apply all-pass transforms to speaker adaptation, we now discuss several practical issues that arise in this endeavor. Most of these concern, in one way or another, parameter estimation procedures.

3.1. Parameter estimation

Let us assume that the parameters specifying a conformal map Q are to be chosen in order to maximize the likelihood of a set of training data. We shall assume that a Gaussian mixture model (GMM) as in (1) and (2) is associated with each state of a HMM. We shall also assume that the covariance matrix D_k is diagonal such that

$$D_k = \mathrm{diag}\{\sigma_{k0}^2\ \sigma_{k1}^2\ \cdots\ \sigma_{k,L-1}^2\}$$

where L is the (original) feature length.

The adaptation of a single mean is achieved by forming the product $\hat{\mu}_k^{(s)} = A^{(s)}\mu_k$ for some speaker-dependent transformation matrix $A^{(s)} = A(\alpha)$. More precisely,

$$\hat{\mu}^{(s)}_{kn} = \sum_{m=0}^{L-1} a_{nm}\mu_{km} \tag{33}$$

for all $n = 0, 1, \ldots, L' - 1$, where the components $\{a_{nm}\}$ of the transformation matrix are given by (24). Thus (1) and (2) must be modified to read

$$P(x; \alpha, \Lambda_k) = \sum_{k=1}^{K} q_k P(x; \alpha, \Lambda_k)$$

where

$$P(x; \alpha, \Lambda_k) = \frac{1}{\sqrt{|2\pi D_k|}} \exp\left[-\frac{1}{2}(x - A^{(s)}\mu_k)^{\mathrm{T}} \times D_k^{-1}(x - A^{(s)}\mu_k)\right] \tag{34}$$

Parameter optimization is most easily accomplished through recourse to the EM algorithm. The EM algorithm requires the formulation of an *auxiliary function* (Dempster et al., 1977), which is equivalent to the expected value of the log-likelihood of some set of training data given the current estimate of the model's parameters. Hence, define a set $\mathscr{X}^{(s)} = \{x^{(s)}_t\}$ of training data contributed by a single speaker. Ignoring the dependence on the HMM states, the log-likelihood of this set can be expressed as

$$\log P(\mathscr{X}^{(s)}; \alpha, \Lambda) = \sum_t \log P(x^{(s)}_t; \alpha, \Lambda) = \sum_t \log\left[\sum_k q_k P(x^{(s)}_t; \alpha, \Lambda_k)\right]$$

In (McDonough, 2000), the relevant auxiliary function is shown to reduce to

$$\mathscr{G}(\mathscr{X}^{(s)}; \alpha, \Lambda) = \frac{1}{2}\sum_{k,n} \frac{c^{(s)}_k}{\sigma^2_{kn}} (\tilde{\mu}^{(s)}_{kn} - \hat{\mu}^{(s)}_{kn})^2 \tag{35}$$

where $c^{(s)}_{k,t}$ is the posterior probability that $x^{(s)}_t$ was drawn from $P(x; \alpha, \Lambda_k)$ and

$$\tilde{\mu}^{(s)}_k = \frac{1}{c^{(s)}_k} \sum_t c^{(s)}_{k,t} x^{(s)}_t$$

is the kth speaker-dependent (SD) mean. It is this objective function that is to be *minimized* in the second step of the EM algorithm. As given above, $\mathscr{G}(\alpha) = \mathscr{G}(\mathscr{X}^{(s)}; \alpha, \Lambda)$ represents a continous and continuously differentiable function, and thus is amenable to optimization by any of a number of numerical methods (Luenberger, 1984; Gill et al., 1981). In order to apply such a method, valid expressions for the gradient and (possibly) Hessian of $\mathscr{G}(\alpha)$ must be available. For reasons of brevity, the derivation of such expressions is not included here. The interested reader should see (McDonough, 2000, Section 5.2).

3.1.1. Inclusion of an additive bias

Very often a cepstral mean transformation of the form $\hat{\mu}_k = A\mu_k$ is augmented with an additive bias to model the effect of a channel or any other filtering to which the original speech signal may be subject. This bias is easily incorporated into our prior analysis. Let us define $\breve{\mu}$ as

$$\breve{\mu}_k = \hat{\mu}_k + \hat{b} \tag{36}$$

where $\hat{b}$ is a bias vector whose components are to be estimated along with the other transformation parameters α. Replacing $\hat{\mu}$ with $\breve{\mu}$ in (35) provides

$$\mathscr{G}(\mathscr{X}^{(s)}; A^{(s)}, \Lambda) = \frac{1}{2}\sum_{k,n} \frac{c^{(s)}_k}{\sigma^2_{kn}} \left(\tilde{\mu}^{(s)}_{kn} - \breve{\mu}^{(s)}_{kn}\right)^2 \tag{37}$$

For any given α, it is straightforward to solve for the optimal $\hat{b}$ by taking partials with respect to the components $\hat{b}_n$ on both sides of (37) and equating to zero:

$$\frac{\partial \mathscr{G}}{\partial \hat{b}_n} = -\sum_k \frac{c_k}{\sigma^2_{kn}} [\tilde{\mu}_{kn} - (\hat{\mu}_{kn} + \hat{b}_n)] = 0$$

where the superscript (s) has once more been suppressed. A trivial rearrangement is sufficient to demonstrate that the optimal bias components for a specified α are given by

$$\hat{b}_n(\alpha) = \frac{\sum_k \frac{c_k}{\sigma^2_{kn}} (\tilde{\mu}_{kn} - \hat{\mu}_{kn})}{\sum_k \frac{c_k}{\sigma^2_{kn}}} \tag{38}$$

3.2. Speaker-adapted and incremental training

The optimal APT parameters for a given speaker are determined in part by the current parameters of the relevant HMM. It is equally true, however, that the optimal parameters of an HMM are determined in part by the APT parameters corresponding to the speakers in its training set. Hence, it is necessary to jointly estimate the parameters of the speaker-dependent APTs and speaker-independent HMM. Speaker-Adapted Training (SAT) is an algorithm capable of accomplishing this task (McDonough, 1998).

SAT can be applied to any speaker adaptation scheme based on a linear transformation of the original cepstral means—a property of both APT adaptation as well as the better-known MLLR (Leggetter and Woodland, 1995). When used with the latter, the SAT model is typically initialized with the final, multiple-mixture HMM obtained from conventional training. This approach *cannot* be used in the case of APT adaptation, for the following reason. Due to the highly constrained nature of the transformation, the APT must rotate all cepstral means in a consistent direction if it is to be effective. If a conventionally trained, multiple-mixture HMM is used as a starting point, the offset vectors between the speaker-independent (SI) and speaker-dependent (SD) cepstral means will be essentially *random* due to the training process. Hence, the SD transforms estimated using this initial model will be indistinguishable from the identity, and no improvement in system performance will be achieved. The validity of this argument has been borne out by empirical trials.

Reasonable APT parameters can be estimated by beginning with a conventionally-trained HMM containing a *single* mixture for each state, accumulating speaker-dependent forward-backward statistics, then optimizing the auxiliary function in (35). The determinative factor is not that the HMM is composed of many Gaussian mixture components, but rather that each state is apportioned a single mixture, as this implies each frame in the training set can, in some sense, only be aligned to a single Gaussian density.

After training the single-mixture SAT model as described above, we could simply split the Gaussian densities and continue training, as recommended by (Young et al., 1999). This procedure, however, is very demanding in terms of computational resources. A more efficient solution is provided by the novel single-pass adapted training (SPAT) strategy (McDonough and Byrne, 1999), which is similar in spirit to the single-pass training procedure advocated for use with the Hidden Markov Model Toolkit (HTK) (Young et al., 1999). The complete SPAT procedure is outlined here:

(0) Use the HTK incremental build procedure to obtain conventional, single-mixture (SM) and multiple-mixture (MM), state-clustered SI models.
(1) Perform several iterations of regular SAT beginning with the SM model from Step 0. Retain the SD adaptation parameters for all training speakers.
(2) Beginning with the final, MM model from Step 0, do a forward-backward pass on all utterances in the training set and dump SD statistics. Note that no speaker adaptation is performed on the SI model prior to forward–backward alignment.
(3) Using the SD adaptation parameters from Step 1 and the SD forward–backward statistics from Step 2, perform a regular SAT combination step.
(4) Perform several additional iterations of normal SAT beginning with the model obtained from Step 3 and SD adaptation parameters from Step 1.

In a series of experiments using speech material from the Switchboard Corpus, the model trained using the SPAT procedure performed at least as well as that obtained using a “naive” train and split SAT procedure. This fact together with its more modest computational requirements makes SPAT an appealing choice.

It should be noted that SPAT admits several useful variations; in particular, it is often desireable to estimate several transformations for a single speaker in order to capture fine inter-speaker variations. This can be achieved by partitioning all Gaussian components in an HMM into several regression classes and estimating a unique speaker-dependent transformation for each. In a straightforward modification of SPAT, regression classes are added incrementally to the single-mixture model by splitting the existing regression classes using a K-means-like procedure, and then performing several iterations of conventional speaker-adapted training. Thereafter, the new regression classes can be "transferred" to the multi-mixture model. An exact description of this procedure can be found in (McDonough, 2000, Section 6.2).

4. Speech recognition experiments

In this section we summarize the speech recognition experiments undertaken to illustrate the effectiveness APT-based speaker adaptation, and to compare it with the popular MLLR technique. Two corpora were used for these experiments: the Switchboard Corpus and the English Spontaneous Scheduling Task (ESST).

4.1. Switchboard experiments

The Switchboard Corpus is a collection of approximately 2500 conversations conducted over standard US telephone lines between persons previously unknown to each other. This corpus abounds in all the phenomena that make the automatic recognition of spontaneous speech a difficult task: extreme co-articulation effects, stops and restarts, ungrammatical word usage, and vowel reduction comprise a partial list.

Of the complete Switchboard Corpus, approximately 140 h of data are set aside for system training. For the purpose of the experiments described below, however, a subset of the complete training corpus was identified. This training set, dubbed *MsTrain*, is composed of nearly 800 complete conversations spoken by 409 speakers, and totals 50.0 h of speech. The test set used in all Switchboard was comprised of both sides of 19 Switchboard conversations, for a total of 18,000 words.

The features used for speech recognition were composed of the first 12 perceptual linear prediction (PLP) cepstral coefficients (Hermansky, 1990) along with first and second order difference coefficients derived from these. Parameters corresponding to short-time energy and its first and second order difference were also estimated, for a total feature length of 39. Cepstral mean subtraction was applied to the features of the test and training sets on a per conversation side basis; no other feature normalization was used.

For experiments on the Switchoard Corpus, all HMM training and test was conducted using HTK (Young et al., 1999), which was augmented with the Homewood Extensions (McDonough, 1999). The HMMs were trained with cross-word triphones. Each triphone was composed of three states, and each state was composed of 12 Gaussians. The standard HTK implementation of the decision tree algorithm was used to generated the state clusters of the HMM; the total number of HMM state clusters used with the MsTrain set was 6,712. All word-error rates tabulated below were generated by rescoring a set of trigram lattices with a modified version of the HKT tool `HVite`. The vocabulary used in generating and rescoring the lattices contained approximately 40,000 words.

4.1.1. Rapid adaptation

The results of a set of experiments conducted to compare full-matrix MLLR and APT-based adaptation on a task with limited enrollment data are given in Table 1; in keeping with popular usage, we will hereafter refer to this scenario as *rapid adaptation*. For these experiments, one global transformation was used for each speaker and cepstral mean subtraction (CMS) was applied on a per utterance basis. All systems were trained on the MsTrain set; the SPAT and basic SAT procedures were used for the APT- and MLLR-based systems, respectively. The errorful transcripts used for unsupervised parameter estimation, be it MLLR or APT, were obtained with the unadapted baseline system, which achieved a WER of 41.5%. Either one or nine free parameters were used to specify the RAPT and SLAPT transforms, as indicated by the "–1" and "–9" suffixes on the column headings. As is apparent from the table, when 2.5 min of data were used during the unsupervised estimation of transformation parameters, the performance of MLLR and the nine-parameter APT system were nearly identical. As the amount of adaptation data was reduced, however, the performance of the MLLR system quickly deteriorated, suffering a catastrophic degradation at 10.0

Table 1
Results of unsupervised rapid adaptation experiments for systems trained on the MsTrain set

Enrollment set	% Word error rate				
	RAPT-1	RAPT-9	SLAPT-1	SLAPT-9	MLLR
Baseline	41.5				
2.5 min	38.5	37.3	38.4	37.4	37.1
60 s	38.3	37.4	38.2	37.5	37.5
30 s	38.5	37.6	38.3	37.7	37.9
10 s	38.7	37.8	38.6	38.0	40.1
5 s	38.8	37.9	38.6	38.2	45.5

s and less. The APT-based system, on the other hand, experienced only marginal performance degradation, providing a reduction in WER of 3.5% absolute with only 5.0 s of enrollment data. This difference in characteristics is surely due to the sparse parameterization of the APT.

4.1.2. Multi-regression class adaptation

Another set of speaker adaptation experiments were undertaken to compare the effectiveness of APT-based adaptation to MLLR when both are applied with multiple regression classes. The systems used to obtain the WER results given in Table 2 were trained on the MsTrain set with per-conversation side CMS, but no VTLN. The initial recognition pass was conducted with an unadapted SI acoustic model and a bigram language model, yielding a WER of 40.6%. The one-best hypotheses from the initial pass were then used to perform unsupervised parameter estimation. The subsequent recognition passes used speaker adaptation as well as a trigram language model. In each, the RAPT was augmented with an additive bias component applied to the static cepstral features but not to the deltas nor delta–deltas; the HMM used in recognition was trained with the incremental procedure described in (McDonough, 2000, Section 6.2). As shown in the table, the number of regression classes was varied. Clearly, the use of more regression classes to capture fine inter-speaker differences results in ever increasing reductions in WER. The best system apportioned $528 = 24 \times (9 + 13)$ total parameters to each speaker and achieved a WER of 35.6%.

A second set of experiments was conducted to compare the performance of APT-based adaptation to that of the popular MLLR; the results of these experiments are shown in Table 3. In this case, the transform was composed of a full, unconstrained matrix augmented with an additive bias term applied to the entire cepstral feature, including deltas and delta-deltas. These systems were trained on the MsTrain set with the basic SAT procedure (Anastasakos et al., 1996). As shown in the table, the best system achieved a WER of 36.3%, which is significantly worse than that obtained with the best APT-based system. Moreover, the use of two regression classes provided no significant reduction in WER over the single-class model, which apportioned $1560 = 1 \times (39 \times 40)$ transform parameters to each speaker. This is not surprising given the unsupervised nature of the adaptation, the high initial word error rate, and the large number of parameters present in each individual MLLR matrix. That the opposite trend was observed for the APT is a consequence of its parsimonious parameterization.

Table 2
Word error rate results of unsupervised speaker adaptation experiments using the for systems trained on the MsTrain set without VTLN

No. regression classes	% Word error rate	
	RAPT-1	RAPT-9
Baseline	40.6	
1	38.2	37.3
2		37.0
4		36.3
8		36.1
16		36.1
24		35.6

Table 3
MLLR/SAT results obtained for systems trained on the MsTrain set without VTLN

No. regression classes	% Word error rate
Baseline	40.6
1	36.9
2	36.3
4	37.3

4.2. English spontaneous scheduling task experiments

The speech experiments described below were conducted with the Janus Recognition Toolkit (JRTk), which is maintained and developed jointly at Universität Karlsruhe, in Karlsruhe, Germany and at the Carnegie Mellon University in Pittsburgh, PA, USA. In the recent past, the Homewood Extensions (McDonough, 1999) were ported to JRTk by the first author. Moreover, their capabilities were enhanced to support SAT using a maximum mutual information criterion (McDonough, 2001).

Training was conducted on the English Spontaneous Scheduling Task (ESST), which contains approximately 35 h of speech contributed by 242 speakers. ESST is also a conversational LVCSR task, in which participants discuss travel arrangements and schedule business meetings. As such, it contains the same conversational speech artifacts seen in Switchboard. Unlike Switchboard, however, this data was collected with Sennheiser close-talking microphones instead of standard telephones. For these experiments, we used a baseline model with 48 Gaussians for each of 2339 codebooks. The ESST test set contains 22,889 total words.

All speech data was digitally sampled at a rate of 16 kHz. The speech features used for all experiments were obtained by estimating 13 cepstral components, along with their first and second differences. Features were calculated every 10 ms using a 16 ms sliding window. Speaker-dependent frequency-domain vocal tract length normalization (VTLN) was used in calculating all speech feaures for both training and test.

Unsupervised speaker adaptation for all test conditions requiring it, was performed on the errorful test set transcriptions obtained with the unadapted, baseline recognizer. MLLR and APT parameter estimation was conducted by iterating twice over each conversation side in the test set, which implies that approximately four minutes of speech per speaker was available for unsupervised adaptation. The results of the ESST experiments are summarized in Table 4.

Table 4
Word error rate results comparing SLAPT- and MLLR-adaptation on the ESST corpus with VTLN prior to adaptation

No. regression classes	% Word error rate	
	SLAPT-9	MLLR
Baseline	27.3	
1	24.66	24.03
2	24.03	23.78
4	23.67	24.49
8	23.28	25.15
12	23.06	N/A
16	22.61	N/A
24	22.40	N/A

In the SLAPT-based systems, the means of the speaker-independent HMM were *extended* from their original length of 39 to a final length of 78 during SATraining. This was accomplished by exploiting the fact that the summation in (25) is *infinite*. Hence, a finite number of SLAPT parameters induce a transformation matrix that is arbitrarily "wide;" it need not be truncated after only 13 columns.

Several things are apparent upon examining Table 4. First, the best SLAPT result (22.40%) is more than a full point better than the best MLLR result (23.78%). Second, SLAPT adaptation is still improving with the addition of regression classes, while MLLR quickly peaks with only two regression classes. Third, the WER reductions afforded by SLAPT-based adaptation are *additive* with those given by VTLN.

5. Conclusions and future work

In this work we have described a class of conformal maps known as all-pass transforms

(APTs). Forming the composition of the *z*-transform of a cepstral sequence with an APT, it is possible to obtain a transformed cepstral sequence using only linear operations on the original cepstra. Moreover, the result of this transformation in the spectral domain can be equated to a warping or rescaling of the frequency axis similar to that seen in conventional vocal tract length normalization (VTLN). The transformation effected by an APT is more general than VTLN, however, and can be made arbitrarily complex by increasing the number of free parameters specifying the map.

In a set of unsupervised speaker adaptation experiments conducted on conversational speech material from the Switchboard Corpus, we have demonstrated APT-based adaptation is more effective than MLLR on tasks involving 30 s or less of unsupervised enrollment data. A second set of speech recognition experiments were conducted on speech material extracted from the English Spontaneous Scheduling Task. Here the unsupervised adaptation task involved several minutes of enrollment data; hence, multiple regression classes could be effectively used for both APT- and MLLR-adaptation. Once more, APT adaptation, with a final WER of 22.40%, proved more effective than MLLR, which achieved a WER of 23.78%. What is even more compelling is that these results were obtained with cepstral features to which VTLN had already been applied; i.e., the WER reductions afforded by APT adaptation are additive with those provided by VTLN.

In as yet unpublished work (McDonough, 2003), we made similar comparisons of APT adaptation and MLLR on the Switchboard Corpus, both without and with VTLN. We found that without VTLN, the best MLLR and APT systems achieved word error rates (WERs) of 43.0% and 40.2% respectively. Similarly, with VTLN the respective error rates were 40.3%, and 39.2%, so that APT adaptation is significantly better in both cases. Further effort must be devoted to more comparisons of APT adaptation and MLLR, as well as to their *combination*. More work is also required to refine the current training procedure. In particular, the incremental addition of regression classes and transform parameters is an area requiring more study.

Most current state-of-the-art systems apply one or more linear transformations (e.g., linear discriminant analysis (Kumar and Andreou, 1998), diagonalizing transforms (Saon et al., 2000)) to the raw cepstral features before using them for recognition. APT adaptation, on the other hand, assumes that the raw cepstral features are available, as it is formulated so as to exploit the special characteristics of such features. In (McDonough, 2003), we also report the results of some initial investigations into the use of APT adaptation in situations where one or more linear transformations are applied to the raw cepstra. In particular, we combined APT adaptation with the linear feature transformation inherent in the estimation of semi-tied covariance (STC) matrices (Gales, 1999). We found that with a single APT transformation per speaker, the application of STC reduced the WER from 42.9% to 39.4%.

Recently there has been a renewed interest in discriminative training techniques such as maximum mutual information (MMI) parameter estimation (Woodland and Povey, 2000). MMI parameter estimation has already been successfully combined with SAT for the case of MLLR adaptation (McDonough, 2001). It would also be of interest to apply MMI-SAT when APT- as opposed to MLLR-based adaptation is used.

Acknowledgements

The first author would like to express his gratitude to Andreas Andreou, William Byrne, Pablo Iglesias, Frederick Jelinek, and Steven Zelditch for reading an earlier version of this work, and for providing many helpful comments and suggestions.

References

Acero, A., 1990. Acoustical and environmental robustness in automatic speech recognition. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Anastasakos, T., McDonough, J., Schwartz, R., Makhoul, J., 1996. A compact model for speaker-adaptive training. In: Proc. ICSLP.

Andreou, A., Kamm, T., Cohen, J., 1994. Experiments in vocal tract normalization. In: Proc. CAIP Workshop: Frontiers in Speech Recognition II.

Bocchieri, E., Digalakis, V., Corduneanu, A., Boulis, C., 1999. Correlation modeling of MLLR transform biases for rapid HMM adaptation to new speakers. In: Proc. ICASSP, Vol. II, pp. 773–776.

Churchill, R.V., Brown, J.W., 1990. Complex Variables and Applications, fifth ed. McGraw-Hill, New York.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. 39B, 1–38.

Digalakis, V., Rtischev, D., Neumeyer, L., 1995. Fast speaker adaptation using constrained estimation of gaussian mixtures. IEEE Trans. Speech Audio Process. 3, 357–366.

Digalakis, V., Berkowitz, S., Bocchieri, E., Boulis, C., Byrne, W., Collier, H., Corduneanu, A., Kannan, A., Khudanpur, S., Sankar, A., 1996. Rapid speech recognizer adaptation to new speakers. In: Proc. ICASSP, Vol. I, pp. 339–341.

Ding, G.-H., Zhu, Y.-F., Li, C., Xu, B., 2002. Implementing vocal tract length normalization in the MLLR framework. In: ICSLP, pp. 1389–1392.

Eide, E., Gish, H., 1996. A parametric approach to vocal tract length normalization. In: Proc. ICASSP, Vol. I, pp. 346–348.

Gales, M.J.F., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech Language 12, 75–98.

Gales, M.J.F., 1999. Semi-tied covariance matrices for hidden markov models. IEEE Trans. Speech Audio Process. 7, 272–281.

Gales, M.J.F., Woodland, P.C., 1996. Mean and variance adaptation within the MLLR framework. Computer Speech Language 10, 249–264.

Gill, P.E., Murray, W., Wright, M.H., 1981. Practical Optimization. Academic Press, London.

Gunawardana, A., Byrne, W., 2000. Robust estimation for rapid speaker adaptation using discounted likelihood techniques, IEEE ICASSP, 5–9 June 2000, Vol. 2, pp. II985–II988.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 87 (4), 1738–1752.

Kannan, A., Khudanpur, S., 1996. Tree-structured models of parameter dependence for rapid adaptation in large vocabulary conversational speech recognition. In: Proc. ICASSP, Vol. II, pp. 769–772.

Kuhn, R., Junqua, J.-C., Nguyen, P., Niedzielski, N., 2000. Speaker adaptation in eigenvoice space. IEEE Trans. Speech Audio Process. 8 (6), 695–707.

Kumar, N., Andreou, A.G., 1998. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. Speech Commun. 26, 238–297.

Lee, L., Rose, R.C., 1996. Speaker normalization using efficient frequency warping procedures. In: Proc. ICASSP, Vol. I, pp. 353–356.

Leggetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech Language 9, 171–185.

Luenberger, D.G., 1984. Linear and Nonlinear Programming, second ed. Addison-Wesley, New York.

Masry, E., Stieglitz, K., Liu, B., 1968. Bases in hilbert space related to the representation of stationary operators. SIAM J. Appl. Math. 16, 552–562.

McDonough, J.W., 1998. On the estimation of optimal regression classes for speaker adaptation. Tech. Rep. 36, Center for Language and Speech Processing, The Johns Hopkins University.

McDonough, J.W., 1999. The Homewood extensions. Tech. Rep. 39, Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD.

McDonough, J.W., 2000. Speaker compensation with all-pass transforms. Ph.D. thesis, The Johns Hopkins University, Baltimore, MD.

McDonough, J.W., 2001. On maximum mutual information speaker-adapted training. Tech. Rep. 103, Universität Karlsruhe.

McDonough, J., Waibel, A., 2003. Performance comparisons of all-pass transform adaptation with maximum likelihood linear regression. Tech. Rep. 102, Universität Karlsruhe.

McDonough, J.W., Byrne, W., 1999. Single-pass adapted training with all-pass transforms. In: Proc. Eurospeech.

McDonough, J., Byrne, W., Luo, X., 1998. Speaker normalization with all-pass transforms. In: Proc. ICSLP.

Oppenheim, A.V., Johnson, D.H., 1972. Discrete-time representation of signals. Proc. IEEE 60 (6), 681–691.

Oppenheim, A.V., Schafer, R.W., 1989. Discrete-Time Signal Processing. Prentice-Hall, Englewood Cliffs, NJ.

Pitz, M., Molau, S., Schlüter, R., Ney, H., 2001. Vocal tract normalization equals linear transformation in cepstral space. In: Eurospeech, pp. 721–724.

Pye, D., Woodland, P.C., 1997. Experiments in speaker normalisation and adaptation for large vocabulary speech recognition. In: Proc. ICASSP, Vol. II, pp. 1047–1050.

Sankar, A., Lee, C.-H., 1996. A maximum-likelihood approach to stochastic matching for robust speech recognition. IEEE Trans. Speech Audio Process. 4 (3), 190–201.

Saon, G., Zweig, G., Padmanabhan, M., 2000. Maximum likelihood discriminant feature spaces. In: Proc. ICASSP.

Shikano, K., 1986. Evaluation of LPC spectral matching measures for phonetic unit recognition. Tech. Rep., Computer Science Department, Carnegie Mellon University, Pittsburgh, PA.

Wegmann, S., McAllaster, D., Orloff, J., Peskin, B., 1996. Speaker normalization on conversational telephone speech. In: Proc. ICASSP, Vol. I, pp. 339–341.

Woodland, P., Povey, D., 2000. Large scale discriminative training for speech recognition. In: ISCA ITRW Automatic

Speech Recognition: Challenges for the Millenium, pp. 7–16.

Young, S., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 1999. The HTK Book. Entropic Software, Cambridge.

Zue, V., 1971. Translation of divers' speech using digital frequency warping. Tech. Rep. 101, Res. Lab. Eltron., Massachusetts Institute of Technology, Cambridge, MA.