SPEAKER NORMALIZATION WITH ALL-PASS TRANSFORMS

John McDonough *William Byrne* *Xiaoqiang Luo*

Center for Language and Speech Processing
The Johns Hopkins University
Baltimore, MD, USA 21218
e-mail: jmcd@mail.clsp.jhu.edu

ABSTRACT

Speaker normalization is a process in which the short-time features of speech from a given speaker are transformed so as to better match some speaker independent model. Vocal tract length normalization (VTLN) is a popular speaker normalization scheme wherein the frequency axis of the short-time spectrum associated with a speaker's speech is rescaled or warped prior to the extraction of cepstral features. In this work, we develop a novel speaker normalization scheme 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 propose a class of such maps, designated all-pass transforms for reasons given hereafter, and in a set of speech recognition experiments conducted on the Switchboard Corpus demonstrate their capacity to achieve word error rate reductions of 3.7% absolute.

1. INTRODUCTION

In *speaker normalization*, we attempt to transform the short-time features of a given speaker's speech in such a fashion that they will better match a speaker independent model. This normalization is performed with the intention of reducing the word error rate of a large vocabulary continuous speech recognition system.

Undoubtedly one of the most popular speaker normalization schemes is vocal tract length normalization (VTLN), a technique which has enjoyed a large coverage in the literature [10]. In a typical implementation of VTLN, a digitally-sampled utterance is windowed to isolate a short segment, then analyzed with the FFT to obtain the short-time spectrum. This spectrum may or may not be subjected to smoothing via the estimation of a linear predictive model. Normalization is achieved by warping the frequency axis of the short-time spectrum using a suitable parameterized function; the parameter values are estimated individually for each speaker. The normalized cepstra serving as features for a speech recognition system are then extracted by taking the inverse FFT on the warped spectrum.

In the present work, we extend and generalize the speaker normalization paradigm outlined above. Our point of departure is the observation that the parameterized warp function used in most VTLN implementations can be approximated to a reasonable degree by the *bilinear transform* [1, 6]. The bilinear transform is a complex-valued function of a single complex argument, is uniquely specified by a single parameter, and is analytic in an annular region including the unit circle. As discussed in prior work [4, 6, 7], the latter property implies that a sequence of transformed cepstra can be obtained through a *linear* transformation of a sequence of initial cepstra. In addition, the bilinear transform maps the unit circle in the complex plane back onto the unit circle, and for this reason is referred to as an *all-pass* transform. As we will shortly demonstrate, it is possible to formulate more general all-pass transforms, which are specified by more than one parameter and thus are potentially more powerful transformations for use in speaker normalization. These transforms, like the bilinear transform, are analytic on the unit circle, and therefore can be implemented through a linear transformation of cepstral features.

The balance of this work is organized as follows. In Section 2., we summarize several important properties of all-pass transforms, and from these construct the theoretical framework necessary to calculate a normalized cepstral sequence from an un-normalized initial sequence. Section 3. discusses a few pertinent details of the parameter estimation which must be performed in implementing a practical speaker normalization scheme. In Section 4. we document the results of several speech recognition experiments conducted to date. These experiments were undertaken to compare conventional VTLN to the speaker normalization paradigm proposed in this work, using both the bilinear transform as well as the more general all-pass transforms. Finally, in Section 5. we summarize the results of our initial experiments, speculate on their meaning, and discuss plans for future work.

2. THEORETICAL DEVELOPMENT

Consider a real, even cepstral sequence $c[n]$ and its associated z-transform $C(z)$, here expressed as

$$C(z) = \sum_{n=-\infty}^{\infty} c[n]\, z^n \tag{1}$$

With this definition $c[n]$ can be recovered from $C(z)$ through the contour integral

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

for all $n = 0, \pm 1, \pm 2, \ldots$. In what follows, we shall consider Equations (1–2) as comprising the *transform pair* $c[n] \leftrightarrow C(z)$.

Consider now a conformal map $Q(z)$, which we hope to use as a mechanism for calculating a normalized cepstral sequence $\hat{c}[n]$ from the initial sequence $c[n]$. The bilinear transform (BLT) is a conformal map well-suited to this application; it can be expressed as

$$Q(z) = \frac{z-\alpha}{1-\alpha z} \tag{3}$$

where α is real and $|\alpha| < 1$. It is also possible to formulate more general conformal maps which subsume the bilinear transform, as indicated by

$$\begin{aligned} Q(z) &= \underbrace{\frac{z-\alpha}{1-\alpha z}}_{} \quad \underbrace{\frac{z-\beta}{1-\beta^* z}\,\frac{z-\beta^*}{1-\beta z}}_{} \quad \underbrace{\frac{1-\gamma^* z}{z-\gamma}\,\frac{1-\gamma z}{z-\gamma^*}}_{} \\ &= \quad A(z) \qquad\qquad B(z) \qquad\qquad\qquad G(z) \end{aligned} \tag{4}$$

where β and γ are complex quantities, such that $\|\beta\|, \|\gamma\| < 1$. The most salient characteristics of either map are that:

1. The unit circle is mapped back to the unit circle, since
$$|Q(e^{j\omega})| = 1 \tag{5}$$
2. The inverse of $Q(z)$ is easily calculated according to
$$Q^{-1}(z) = Q(z^{-1}) \tag{6}$$

Equality (5) is indeed the reason that conformal maps such as (3–4) are generally referred to as all-pass systems in the digital signal processing literature [8, Section 5.5]; such systems have uniform frequency response and thus "pass" signals of all frequencies with neither attenuation nor amplification. Although they are not discussed here, it is possible to devise even more general conformal maps than (4) which still retain these properties [5].

Using an all-pass transform (APT), we should like to transform a cepstral sequence $c[n]$ in some desireable manner. Hence, let us define the z-transform $\hat{C}(z)$ as the composition of $Q(z)$ and $C(z)$, such that $\hat{C}(z) = C(Q(z))$. Furthermore, we should like to associate with $\hat{C}(z)$ a transformed cepstral sequence $\hat{c}[n]$, where $\hat{c}[n] \leftrightarrow \hat{C}(z)$. More formally,

$$\hat{c}[n] = \frac{1}{2\pi j}\oint \hat{C}(z)\, z^{-(n+1)}\, dz \tag{7}$$

$$= \sum_{m=-\infty}^{\infty} c[m] \frac{1}{2\pi j}\oint Q^m(z)\, z^{-(n+1)}\, dz \tag{8}$$

where (7) follows from (8) through use of the series representation (1) for $C(z)$ and subsequent manipulation of the resulting expression. The linearity of the cepstral transformation effected by a conformal map is apparent from (8); this linearity is a direct result of the analyticity of $Q(z)$ on the contour of integration, in this case, the unit circle.

We can exploit the aforementioned analyticity further by forming the transform pair $q[n] \leftrightarrow Q(z)$. For example, it is straightforward to show that $Q(z)$ as given in (3) admits the series representation

$$\begin{aligned} Q(z) &= (z-\alpha)\sum_{n=0}^{\infty} \alpha^n z^n \\ &= -\alpha + (1-\alpha^2)z + \alpha(1-\alpha^2)z^2 + \cdots \end{aligned}$$

From the final equality, the coefficients $q[n]$ of the series expansion are available by inspection. It is also possible to obtain series expansions for $B(z)$ and $G(z)$ appearing in (4), see [5] for details. Thus, upon defining the transform pairs $a[n] \leftrightarrow A(z)$, $b[n] \leftrightarrow B(z)$, and $g[n] \leftrightarrow G(z)$, the final sequence $q[n]$ will be given by

$$q[n] = a[n] * b[n] * g[n] \tag{9}$$

where $*$ is the convolution operator or *Cauchy product* [2, Section 52]. Furthermore, the analyticity of $Q^m(z)$ can be exploited to form a transform pair $q^{(m)}[n] \leftrightarrow Q^m(z)$ for every $m \geq 0$, such that

$$q^{(m)}[n] = \frac{1}{2\pi j}\oint Q^m(z)\, z^{-(n+1)}\, dz \tag{10}$$

In general, the sequences $q^{(m)}[n]$ will have infinite extent for both positive and negative values of n.

From (10) we deduce two things: Firstly, a simple application of the *Cauchy integral formula* [2, Section 39] reveals that $q^{(0)}[n]$ is the unit sample sequence, such that

$$q^{(0)}[n] = \begin{cases} 1; & \text{for } n = 0 \\ 0; & \text{otherwise} \end{cases} \tag{11}$$

Secondly, as $Q^m(z) = Q(z) \times Q^{m-1}(z)$, the several sequences $q^{(m)}[n]$ for all $m > 1$ can be calculated based solely on knowledge of $q^{(1)}[n]$ via the recursion

$$q^{(m)}[n] = q^{(m-1)}[n] * q^{(1)}[n] \tag{12}$$

Hence, comparing (10) with the integral in (8), we discover the desired cepstra are available from

$$\hat{c}[n] = \sum_{m=-\infty}^{\infty} c[m] q^{(m)}[n] \tag{13}$$

As $c[m]$ is even, it is uniquely specified by its causal portion. Following the example set by others [8, Chapter 12], let us make use of this fact to define the sequence $x[n]$ as

$$x[n] = \begin{cases} 0; & n < 0 \\ c[0]; & n = 0 \\ 2c[n]; & n > 0 \end{cases} \tag{14}$$

This latter sequence is the one most often associated with the term *cepstrum.* In this case, $c[n]$ can be recovered from $x[n]$ through the relation

$$c[n] = \tfrac{1}{2}(x[n] + x[-n]) \tag{15}$$

In addition, further consideration of Eqn. (6) reveals that

$$q^{(-m)}[n] = q^{(m)}[-n] \tag{16}$$

If we also define a sequence $\hat{x}[n]$ as the causal portion of $\hat{c}[n]$, and substitute (14–16) into (13), we deduce that it is possible to obtain $\hat{x}[n]$ from

$$\hat{x}[n] = \sum_{m=0}^{\infty} a_{nm}\, x[m] \tag{17}$$

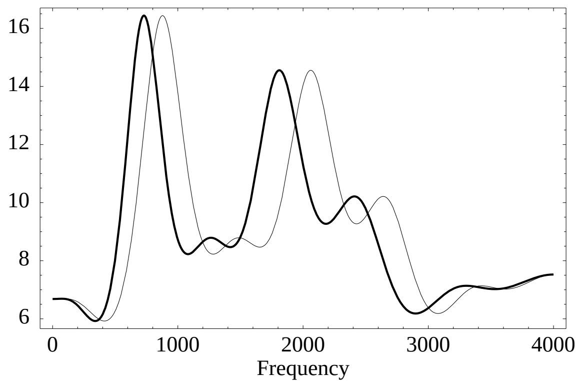

Figure 1. Original (thin line) and transformed (thick line) short-term spectra for a male test speaker regenerated from cepstral coefficients 0–14. The transformed spectrum was produced with the BLT by setting $\alpha = 0.10$.

where

$$a_{nm} = \begin{cases} q^{(m)}[0], & \text{for } n = 0, m \geq 0 \\ 0, & \text{for } n > 0, m = 0 \\ \left(q^{(m)}[n] + q^{(m)}[-n]\right), & \text{for } n, m > 0 \end{cases} \quad (18)$$

are the components of the *transformation matrix* $A = \{a_{nm}\}$.

Figure 1 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 in (17–18); the conformal map used in this case was a bilinear transform with $\alpha = 0.10$. 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.

3. PARAMETER ESTIMATION

Prior to speech recognition, the parameter α must be estimated individually for each speaker in a test or training set. This is accomplished by taking a segment of speech from a given speaker, generating transformed cepstra corresponding to a given warp factor and then calculating the likelihood of the transformed cepstra using a simple Gaussian mixture model (GMM). The likelihood assigned a set $\mathcal{X}^{(s)}$ of features from speaker s can be expressed as

$$\mathcal{L}(\mathcal{X}^{(s)}; \alpha) = \sum_i \log \sum_k q_k P_k(\hat{x}_i^{(s)}(\alpha)) \quad (19)$$

where the probability density function P_k of the k^{th} mixture component is multi-variate normal with diagonal covariance, q_k is the *a priori* probability of the k^{th} component, and $\hat{x}_i^{(s)}(\alpha)$ is the i^{th} transformed feature. The optimal warp parameter α_* is determined from the maximum likelihood criterion:

$$\alpha_* = \operatorname*{argmax}_{\alpha} \mathcal{L}(\mathcal{X}^{(s)}; \alpha) \quad (20)$$

The transformed features $\{\hat{x}_i^{(s)}(\alpha_*)\}$ are subsequently used for speech recognition. As there is no closed form solution for (20), it is necessary to use a numerical search to find α_*. Good results have been obtained with *Brent's method* [9, Section 10.2]. Estimation of optimal parameters for the general all-pass transforms is discussed in [5].

Sankar and Lee [11] point out that if $x_i^{(s)}$ and $\hat{x}_i^{(s)}$ are the original and transformed features respectively, then the log-likelihood of the former is *actually* given by

$$\log P(x_i^{(s)}) = \log J(\alpha) + \log P(\hat{x}_i^{(s)}; \Lambda) \quad (21)$$

where $J(\alpha)$ is the *Jacobian* of the transformation taking $x_i^{(s)}$ to $\hat{x}_i^{(s)}$. The transformation in the present instance is linear such that $\hat{x}_i^{(s)} = Ax_i^{(s)}$, and $J(\alpha)$ reduces to

$$J(\alpha) = \det A(\alpha)$$

Equation (21) implies the actual training set log-likelihood can be expressed as

$$\mathcal{L}'(\mathcal{X}^{(s)}; \alpha) = N \log J(\alpha) + \mathcal{L}(\mathcal{X}^{(s)}; \alpha)$$

where N is the total number of training samples and $\mathcal{L}(\mathcal{X}^{(s)}; \alpha)$ is defined in (19). The actual Jacobian can calculated as the product of the eigenvalues of A; the latter can be determined through use of the *Schur decomposition* [3, §7.1], which is formulated specifically to handle unsymmetric matrices such as A.

4. SPEECH RECOGNITION EXPERIMENTS

The speech recognition experiments discussed below were conducted using training and test material extracted from the *Switchboard Corpus*. Of the complete Switchboard Corpus, approximately 140 hours of data are set aside for system training. In order to obtain fast turnaround, however, a subset of the full training set was identified and used in all speaker normalization experiments. This subset, dubbed *MiniTrain*, is composed of approximately 200 conversations providing a total of 18.6 hours of speech material. Approximately 100 speakers of each gender participate in the MiniTrain conversations. The test set used in all experiments was composed of 19 Switchboard conversations, for a total of 18,000 words.

The features used for speech recognition were composed of mel-frequency cepstral coefficients 1–12 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 42. The mel-frequency cepstral coefficients were calculated using the waveform analysis tools provided with HTK, the Hidden Markov Model Toolkit [12]. Cepstral mean subtraction was applied to the features of the test and training sets on a per utterance basis.

All speech recognition experiments were conducted using a hidden Markov model (HMM) trained with cross-word triphones. Each triphone in the model was composed of three states, and each state was composed of nine Gaussian components. The standard HTK implementation of the decision tree algorithm was used to generated the state clusters of the HMM. The final recognition was composed of approximately 80,000 physical or triphone-level HMMs.

System Description	% Word Error Rate
Baseline	48.9
BLT Test-Only	47.4
BLT Test-Train	45.4
APT Test-Train	45.2

Table 1. Word error rates for lattice rescoring experiments using BLT- and APT-based speaker normalization.

Table 1 provides the results of an initial set of speech recognition experiments conducted to ascertain the effectiveness of bilinear and all-pass transform-based speaker normalization schemes. The results were obtained by rescoring a set of lattices using the appropriate normalization scheme, where the original lattices were generated using the un-normalized or baseline system. After estimation of the appropriate speaker-dependent transformation parameters, the features of the test set were normalized "on the fly" using an appropriately modified version of HVite, the HTK lattice rescoring tool. In the first experiment, the HMM used for rescoring was trained on unnormalized features, but speaker-dependent normalization was applied to the features of the test set. This test condition provided a reduction in word error rate (WER) of 1.5% with respect to the baseline. In the next experiment, BLT normalization was applied to both test and training sets. In order to achieve rapid experimental turnaround, an HMM was trained on the normalized features using the HTK *single-pass training* procedure [12] starting from the conventionally-trained HMM and un-normalized mel-frequency cepstra. Single pass training was followed by an additional four iterations of conventional training using the normalized cepstra. This test condition provided an additional 2.0% reduction in WER with respect to test-only normalization. In the last tabulated experiment, speaker normalization on both test and training sets was performed with a three-parameter APT. This condition provided only a marginal gain over the best WER reduction using the simpler BLT.

5. CONCLUSIONS

In this work, we have investigated the use of the bilinear transform (BLT) as a means of speaker normalization, which is undertaken to improve the performance of a large vocabulary speech recognition system. We have also derived a generalization of the BLT, dubbed the all-pass transform (APT), and compared its performance to that of the BLT. In a set of speech recognition experiments conducted on test and training material abstracted from the Switchboard Corpus, we recorded absolute word error rate (WER) reductions of 3.5% and 3.7% for the BLT and APT respectively beginning from a baseline system with a WER of 48.9% .

A principal result of our theoretical development is that the BLT and APT can be represented as linear transformations in the cepstral domain. The cepstral domain linearity of these transforms implies the Jacobian likelihood normalization factor, whose use is essential for accurate parameter estimation, can be easily calculated for both. This linearity also implies that the transforms in question can be applied to the cepstral *means* of a hidden Markov model (HMM) instead of the cepstral features used for speech recognition, resulting in an instance of speaker *adaptation* as opposed to speaker *normalization*. The use of the BLT and APT in speaker adaptation will be the topic of future work and publications.

REFERENCES

[1] A. Acero. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 1990.

[2] Ruel V. Churchill and James W. Brown. *Complex Variables and Applications*. McGraw-Hill, New York, fifth edition, 1990.

[3] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, third edition, 1996.

[4] John McDonough, George Zavaliagkos, and Herbert Gish. An approach to speaker adaptation based on analytic functions. In *Proc. ICASSP*, volume II, pages 721–724, 1996.

[5] John W. McDonough. Speaker normalization with all-pass transforms. Technical Report No. 28, Center for Language and Speech Processing, The Johns Hopkins University, 1998.

[6] John W. McDonough. Transformation of discrete-time sequences with analytic functions. *IEEE Trans. Speech and Audio Proc.*, submitted for publication.

[7] Alan V. Oppenheim and D. H. Johnson. Discrete representation of signals. *Proc. IEEE*, 60(6), June 1972.

[8] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1989.

[9] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge, second edition, 1992.

[10] D. Pye and P. C. Woodland. Experiments in speaker normalisation and adaptation for large vocabulary speech recognition. In *Proc. ICASSP*, volume II, pages 1047–1050, 1997.

[11] Ananth Sankar and Chin-Hui Lee. A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Trans. Speech and Audio Proc.*, 4(3), May 1996.

[12] Steve Young. *The HTK Book*. Entropic Software, Cambridge, 1997.