Open access • Proceedings Article • DOI:10.1109/ICASSP.1999.759778

Speaker adaptation with all-pass transforms — Source link

John McDonough, William Byrne

Institutions: Johns Hopkins University

Published on: 15 Mar 1999 - International Conference on Acoustics, Speech, and Signal Processing

Topics: Speaker recognition, Speech processing, Speaker diarisation, Word error rate and Normalization (statistics)

SPEAKER ADAPTATION WITH ALL-PASS TRANSFORMS

John McDonough and William Byrne

Center for Language and Speech Processing
The Johns Hopkins University
e-mail: {jmcd,byrne}@mail.clsp.jhu.edu

ABSTRACT

In recent work, a class of transforms were proposed which achieve a remapping of the frequency axis much like conventional vocal tract length normalization. These mappings, known collectively as *all-pass transforms* (APT), were shown to produce substantial improvements in the performance of a large vocabulary speech recognition system when used to normalize incoming speech prior to recognition. In this application, the most advantageous characteristic of the APT was its cepstral-domain linearity; this linearity makes speaker normalization simple to implement, and provides for the robust estimation of the parameters characterizing individual speakers. In the current work, we exploit the APT to develop a *speaker adaptation* scheme in which the cepstral means of a speech recognition model are transformed to better match the speech of a given speaker. In a set of speech recognition experiments conducted on the Switchboard Corpus, we report reductions in word error rate of 3.7% absolute.

1. INTRODUCTION

In *speaker adaptation*, we attempt to transform the cepstral means of a hidden Markov model (HMM) so as to better match the characteristics of some speech from a particular speaker. Speaker adaptation is typically undertaken to reduce the error rate of a large vocabulary conversational speech recognition (LVCSR) system. Certainly one of the most effective speaker adaptation methods is maximum likelihood linear regression (MLLR), wherein a transformation matrix is estimated using some speaker-dependent enrollment data, and then used to transform the cepstral means of a speaker-independent HMM via a straightforward matrix-vector multiply [6].

Speaker normalization is closely related to speaker adaptation, inasmuch as it attempts to transform the short-time *features* of a given speaker's speech so as to better match a speaker independent (SI) model. In prior work, 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 [11]. Secondly, both the BLT and APT can be represented as linear transformations in the cepstral domain [1, 7]. 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 un-normalized cepstra . In addition, the linearity of the transformation lends itself to robust estimation of the requisite speaker dependent transformation parameters, a property not shared by any other current VTLN implemention [8].

In the present work, we attempt to apply the BLT and APT, which have previously proven so useful for speaker normalization, to the task of speaker adaptation. In doing so we shall again exploit the cepstral-domain linearity of these transforms, along with their extremely parsimonious parameterization.

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

This material is based upon work supported by the National Science Foundation under Grant No. #IIS-9732388, and carried out at the 1998 Workshop on Language Engineering, Center for Language and Speech Processing, Johns Hopkins University. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Johns Hopkins University. The authors would like to thank Sanjeev Khudanpur for his assistance with the maximum likelihood linear regression experiments described in this work.

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 [9, 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 [7].

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,

$$\begin{aligned} \hat{c}[n] &= \frac{1}{2\pi j} \oint \hat{C}(z)\, z^{-(n+1)}\, dz && (7) \\ &= \sum_{m=-\infty}^{\infty} c[m] \frac{1}{2\pi j} \oint Q^m(z)\, z^{-(n+1)}\, dz && (8) \end{aligned}$$

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 [7] 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* [3, 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* [3, 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 [9, 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

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

where

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

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

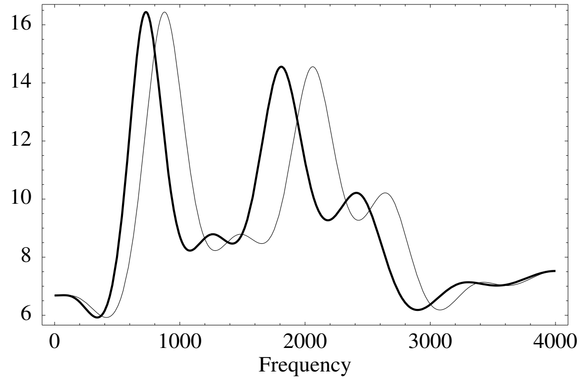

Figure 1: Original (thin line) and transformed (thick line) short-term spectra for a male test speaker regenerated from cepstral coefficients 0–14. The transformed spectrum was produced with the BLT by setting $\alpha = 0.10$.

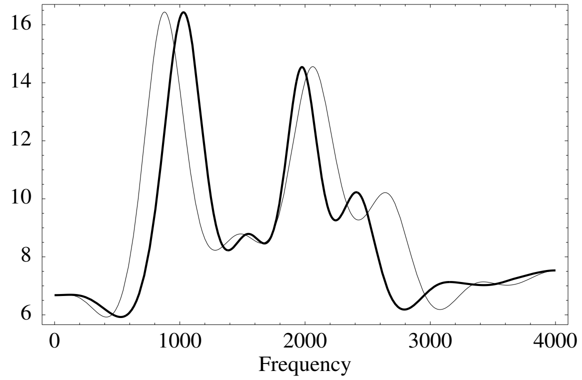

Figure 2: Original (thin line) and transformed (thick line) short-term spectra for a male test speaker regenerated from cepstral coefficients 0–14. The transformed spectrum was produced with the APT.

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.

Shown in Figure 2 are the original and transformed spectra for the same segment of male speech previously plotted in Figure 1. 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 conformal map used in this case was a three-parameter APT with the form given in (4). 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.

3. SPEAKER ADAPTATION

For the purpose of speaker adaptation, we must associate the cepstral sequence $x[n]$ appearing in (17) with the components of the original mean μ_k of an HMM, and the sequence $\hat{x}[n]$ with the components of the transformed mean $\hat{\mu}_k$, such that

$$\hat{\mu}_{kn} = \sum_{m=0}^{L-1} a_{nm}\, \mu_{km}$$

where L is the length of the original mean and the components of the transformation matrix $A = \{a_{nm}\}$ are given in (18). The first- and second-order difference cepstra typically used in LVCSR systems can be transformed in the same way, as they are obtained from appropriate linear combinations of successive cepstral features.

Prior to speech recognition, the transformation parameters α must be estimated individually for each speaker in a test or training set. This is most easily accomplished through recourse to the EM algorithm [4], whose application entails the estimation of an *auxiliary function* and its subsequent maximization with respect to the transform parameters α. Consider a hidden Markov model composed of thousands of individual states; with each state is associated a probability density function composed of several Gaussian components. Let $c_k^{(i)}$ denote the *posterior probability* that the cepstral feature $x^{(i)}$ was drawn from the k^{th} Gaussian component, and let $c_k = \sum_i c_k^{(i)}$ denote the total occupancy count for this component over all frames in a set $\{x^{(i)}\}$ of enrollment data; the several $c_k^{(i)}$ can be calculated via the well-known forward-backward algorithm. Assuming all Gaussian components have diagonal covariance matrices of the form

$$D_k = \text{diag}\left(\sigma^2_{k0}\,,\ \sigma^2_{k1}\,,\ \sigma^2_{k2}\,,\ \ldots\,,\ \sigma^2_{k,L-1}\right)$$

the requisite auxiliary function can be expressed as [7]

$$\mathcal{G}(\alpha) = \sum_k c_k \sum_n \frac{1}{\sigma^2_{kn}} \left(\tilde{\mu}_{kn} - \tfrac{1}{2}\hat{\mu}_{kn}(\alpha)\right) \hat{\mu}_{kn}(\alpha) \qquad (19)$$

where $\tilde{\mu}_k^{(s)} = \left(\sum_i c_k^{(i)} x^{(i)}\right) / c_k$ is the speaker-dependent (SD) mean corresponding to the k^{th} Gaussian component.

As there is no closed form solution providing that $\alpha = \alpha_*$ which achieves a maximum on (19), it is necessary to use a numerical optimization algorithm for this purpose. For the BLT, this optimization devolves to a simple linear search; good results have been obtained with *Brent's method* [10, Section 10.2]. Estimation of optimal parameters for general all-pass transforms can be accomplished with an algorithm based on *conjugate gradients* [10, §10.6] or *Newton's method* [5, §4.4]; expressions for the gradient and Hessian required for either of these methods are developed in [7].

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 adaptation experiments. This subset, dubbed *MiniTrain*, is composed of approximately 200 conversations providing a total of 18.6 hours of speech

	% Word Error Rate	
System Description	0.5 min.	2.5 min.
Baseline	48.9	
BLT Adaptation	45.6	45.5
APT Adaptation	45.2	45.2
GMLLR Adaptation	46.2	45.6

Table 1: Results of lattice rescoring experiments comparing global MLLR to BLT- and APT-based speaker adaptation schemes using either 30 sec. or 2.5 min. of unsupervised enrollment data.

material. Approximately 100 speakers of each gender participate in the MiniTrain conversations. The test set used in all experiments was comprised 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; no other feature normalization was applied.

All speech recognition experiments were conducted using a hidden Markov model 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 model was composed of approximately 3,000 distinct states.

Table 1 provides the results of an initial set of speech recognition experiments conducted to compare the effectiveness of the bilinear and all-pass transform-based speaker adaptation schemes to that of global MLLR adaptation. The results were obtained by rescoring a set of lattices using an appropriately adapted SI model; the lattices were generated using the un-adapted or baseline system. Reported in the right-most column of Table 1 are word error rates for all systems when an entire conversation side—approximately 2.5 minutes of unsupervised enrollment data—was used in estimating the speaker-dependent adaptation parameters, a paradigm typically referred to as *transcription mode*. During parameter estimation, the errorful transcripts obtained with the baseline model were used for the requisite forward-backward passes. As is apparent from the table, the 3.4% absolute word error rate (WER) reduction provided by the BLT was nearly identical to that given by global MLLR. The three-parameter APT scheme gave some additional improvement, for a total WER reduction of 3.7%.

In a second set of tests, only 30.0 sec. of unsupervised enrollment data was used in estimating the adaptation parameters for each test speaker. The results of these tests are given in the left column of Table 1. Due to their extreme parsimony of parameterization, the BLT and APT provided statistically identical performance under both the 30.0 sec. and 2.5 min. test cases. The global MLLR scheme, on the other hand, saw a substantial degradation. The WER reduction provided by global MLLR 30 sec. of enrollment data was 2.7%, while the three-parameter APT gave 3.7% under equivalent conditions.

5. CONCLUSIONS

In this work, we have investigated the use of the bilinear transform (BLT) as the basis of a speaker adaptation scheme, implemented in order to improve the performance of a large vocabulary conversational speech recognition system. We have also presented a generalization of the BLT, known as the all-pass transform (APT), and compared its performance to that of both the BLT and to the well-known maximum likelihood linear regression (MLLR) scheme. Using test and training material abstracted from the Switchboard Corpus, we conducted a set of speech experiments in which unsupervised adaptation was performed on a speaker-independent model. These experiments indicated that the performance of the APT-based speaker adaptation schemes was comparable or better to that of global MLLR when 2.5 min. of unsupervised enrollment data was used for parameter estimation. When the enrollment data was reduced to 30 sec., the APT-based scheme was substantially better than GMLLR; the respective word error rate reductions in this case were 3.7% and 2.7% beginning from an unadapted baseline of 48.9%. Future work will investigate the effect of combining the speaker adaptation schemes proposed here with other speaker compensation techniques such as conventional vocal tract length normalization [11] and speaker-adapted training [2].

6. REFERENCES

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

[2] A. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker-adaptive training. In *Proc. ICSLP*, 1996.

[3] R. V. Churchill and J. W. Brown. *Complex Variables and Applications*. McGraw-Hill, New York, fifth edition, 1990.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihoood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39 B:1–38, 1977.

[5] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, London, 1981.

[6] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech and Language*, pages 171–185, 1995.

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

[8] J. McDonough, W. Byrne, and X. Luo. Speaker normalization with all-pass transforms. In *Proc. ICSLP*, 1998.

[9] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1989.

[10] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, second edition, 1992.

[11] 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.

[12] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK Book*. Entropic Software, Cambridge, 1997.