SPEAKER COMPENSATION WITH SINE-LOG ALL-PASS TRANSFORMS

John McDonough, Florian Metze, Hagen Soltau, and Alex Waibel

Interactive Systems Laboratories
Universität Karlsruhe
Am Fasanengarten 5
76131 Karlsruhe, Germany
{jmcd,metze,soltau,waibel}@ira.uka.de

ABSTRACT

In recent work, we proposed the *rational all-pass transform* (RAPT) as the basis of a speaker adaptation scheme intended for use with a large vocabulary speech recognition system. It was shown that RAPT-based adaptation reduces to a linear transformation of cepstral means, much like the better known maximum likelihood linear regression (MLLR). In a set of speech recognition experiments conducted on the Switchboard Corpus, we obtained a word error rate (WER) of 37.9% using RAPT adaptation, a significant improvement over the 39.5% WER achieved with MLLR. In the present work, we propose the *sine-log all-pass transform* (SLAPT) as a replacement for the RAPT. Our findings indicate the SLAPT is just as effective as the RAPT at reducing WER when used as the basis for a variety of speaker compensation schemes, but in addition conduces to far more tractable computation of transformed cepstral sequences, and the estimation of optimal transform parameters.

1. INTRODUCTION

In prior work we proposed the *rational all-pass transform* (RAPT) as the basis for a variety of practical speaker compensation schemes intended to reduce the word error rate of a large vocabulary speech recognition system. In [6] we considered the use of the RAPT to transform a set of cepstral features to better match a speaker-independent (SI) model, as is typically done in vocal tract length normalization [2]. In [5] we made use of the RAPT to transform the cepstral means of a SI model, as is currently done in most speaker adaptation schemes [4]. We extended this speaker adaptation approach in [8] to handle the case of multiple regression classes; our findings there indicated that RAPT-based adaptation was at least as effective as the better-known maximum likelihood linear regression (MLLR) [4] for moderate amounts of unsupervised enrollment data, but far superior when used with 10 sec. or less of enrollment.

As reported in the aforementioned work, the principal advantages of the RAPT in formulating speaker compensation schemes are its extremely parsimonious parameterization, and its linearity in the cepstral domain; the latter property allows the RAPT to be used as a direct replacement for MLLR in speaker-adapted training (SAT) [8]. The principal *disadvantage* of the RAPT as presented heretofore, is the complexity of the expressions that must be evaluated in calculating transformed cepstral sequences and in performing parameter optimization with respect to a maximum likelihood (ML) criterion. In the present work, we remedy this deficiency by proposing a new functional form for the all-pass transform, dubbed the *sine-log all-pass transform* (SLAPT). The SLAPT retains the very desirable cepstral domain linearity of the RAPT, and is similar in other regards, but in addition is much more tractable when used for the calculation of transformed cepstral sequences and ML parameter estimation. Unlike the RAPT, the SLAPT does not have a rational form. Luckily, however, this is not a drawback for the speaker compensation application which comprises our chief interest.

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. All authors would like to thank the Center for Language and Speech Processing at the Johns Hopkins University for making available the computer resources required to conduct the experiments described in this work.

2. RATIONAL ALL-PASS TRANSFORMS: A REVIEW

Let us begin by reviewing earlier work on the application of all-pass transforms to speaker compensation. 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* $c \leftrightarrow C$.

Consider now a conformal map Q, which we hope to use as a mechanism for calculating a normalized cepstral sequence $\hat{c}$ from the initial sequence c. The bilinear transform (BLT) [1] is a conformal map well-suited to this application; it can be expressed as $Q(z) = (z - \alpha)/(1 - \alpha z)$ 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

$$\begin{aligned} Q(z) &= \underbrace{\frac{z-\alpha}{1-\alpha z}}_{A(z)} \quad \underbrace{\frac{z-\beta}{1-\beta^* z}\,\frac{z-\beta^*}{1-\beta z}}_{B(z)} \quad \underbrace{\frac{1-\gamma^* z}{z-\gamma}\,\frac{1-\gamma z}{z-\gamma^*}}_{G(z)} \end{aligned} \tag{3}$$

where β and γ are complex quantities, such that $|\beta|, |\gamma| < 1$. We shall refer to such functions as *rational* all-pass transforms due to the rational form of (3). The most salient characteristics of these maps are:

$$|Q(e^{j\omega})| = 1 \tag{4}$$

$$Q^{-1}(z) = Q(z^{-1}) \tag{5}$$

Equality (4) is indeed the reason that conformal maps such as (3) are generally referred to as *all-pass systems* in the digital signal processing literature [9, Section 5.5].

Using an RAPT, we should like to transform a cepstral sequence c in some desireable manner. Hence, let us define the composition $\hat{C} = C \circ Q$ and associate with $\hat{C}$ a transformed cepstral sequence $\hat{c}$, such that $\hat{c} \leftrightarrow \hat{C}$. It is straightforward to demonstrate [7, §2.3] that the coefficients of $\hat{c}$ are given by

$$\hat{c}[n] = \sum_{m=-\infty}^{\infty} c[m] \frac{1}{2\pi j} \oint Q^m(z)\, z^{-(n+1)}\, dz \tag{6}$$

The linearity of the cepstral transformation effected by a conformal map is apparent from (6); this linearity is a direct result of the analyticity of Q on the contour of integration, in this case, the unit circle.

We can exploit the aforementioned analyticity further by forming the transform pair $q \leftrightarrow Q$. For example, it is straightforward to show that the simple BLT admits the series representation

$$Q(z) = -\alpha + (1-\alpha^2)z + \alpha(1-\alpha^2)z^2 + \cdots$$

from which the coefficients q of the series expansion are available by inspection. It is also possible to obtain series expansions for B and G appearing in (3); see [7, Appendix C]. Upon defining the transform pairs $a \leftrightarrow A$, $b \leftrightarrow B$, and $g \leftrightarrow G$, the final sequence q for the more general RAPT is readily obtained from

$$q = a * b * g \tag{7}$$

where $*$ is the convolution operator. The analyticity of Q^m can be exploited to form a transform pair $q^{(m)} \leftrightarrow Q^m$ for every $m \geq 0$. In general, the sequences $q^{(m)}$ will have infinite extent for both positive and negative values of n. As $Q^m = Q \times Q^{m-1}$, the several sequences $q^{(m)}$ for all $m > 1$ can be calculated based solely on knowledge of $q^{(1)}$ via the recursion

$$q^{(m)} = q^{(m-1)} * q^{(1)} \tag{8}$$

Moreover, the desired transformed cepstra can be calculated from

$$\hat{c}[n] = \sum_{m=-\infty}^{\infty} c[m] q^{(m)}[n] \tag{9}$$

As c is even, it is uniquely specified by its causal portion. Let us follow [9, Chapter 12] and define $\hat{x}$ as the causal portion of $\hat{c}$, which implies that $\hat{x}$ can be obtained from

$$\hat{x}[n] = \sum_{m=0}^{\infty} a_{nm}\, x[m] \tag{10}$$

where

$$a_{nm} = \begin{cases} q^{(m)}[0], & \text{for } n = 0, m \geq 0 \\ 0, & \text{for } n > 0, m = 0 \\ \left(q^{(m)}[n] + q^{(m)}[-n]\right), & \text{for } n, m > 0 \end{cases} \tag{11}$$

are the components of the *transformation matrix* $A = \{a_{nm}\}$; see [7, §3.3].

3. SINE-LOG ALL-PASS TRANSFORMS

In this section we consider a different type of all-pass transform that shares many of the characteristics of the RAPT, but is simpler in form and thus more amenable to numerical computation. Let us begin by defining the *sine-log all-pass transform* (SLAPT) as

$$Q(z) = z \, \exp\, F(z) \tag{12}$$

where

$$F(z) = \sum_{k=1}^{K} \alpha_k \, F_k(z) \;\text{ for } \alpha_1, \ldots, \alpha_K \in R, \tag{13}$$

$$F_k(z) = j\,\pi \sin\left(\frac{k}{j}\log z\right) \tag{14}$$

and K is the number of free parameters in the transform. The designation sine-log is due to the functional form of F_k. Observe that $F_k(z)$ is single-valued even though $\log z$ is multiple-valued [7, §3.5]. Moreover, applying the well-known relation $\sin z = (e^{jz} - e^{-jz})/2j$ to (14) provides

$$F_k(z) = \frac{\pi}{2}\left(z^k - z^{-k}\right) \tag{15}$$

which is a more tractable form for computation. Note that Q as defined in (12) satisfies (4–5).

In order to calculate the coefficients of a transformed cepstral sequence in the manner described in Section 2, it is first necessary to calculate the coefficients q in the Laurent series expansion of Q; this we do as follows: For F as in (13) set

$$G(z) = \exp F(z) \tag{16}$$

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

$$g[n] = \frac{1}{2\pi j} \oint G(z)\, z^{-(n+1)}\, dz \tag{17}$$

where the contour of integration is the unit circle. The natural exponential admits the series expansion $e^z = \sum_{m=0}^{\infty} \frac{z^m}{m!}$ for all $z \in C$, so that

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

for all $z \in C \backslash \{0\}$. As explained in [7, §3.5] substituting the latter into (17) provides

$$g[n] = \sum_{m=0}^{\infty} \frac{1}{m!} \frac{1}{2\pi j} \oint F^m(z)\, z^{-(n+1)}\, dz \tag{18}$$

The sequence f of coefficients in the series expansion of F is available by inspection from (13) and (15). Defining $f^{(m)}$ such that $f^{(m)} \leftrightarrow F^m$ and applying this definition to (18) we find

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

Moreover, from the Cauchy product it follows $f^{(m)} = f * f^{(m-1)}$ for $m = 1, 2, 3, \ldots$. Equation (16) implies that $Q(z) = z\,G(z)$, so the desired coefficients are given by

$$q[n] = g[n-1] \tag{19}$$

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

The development of this section indicates the primary advantage of the sine-log APT with respect to the rational APT considered in earlier work; i.e., the computations are much simpler: The basic series expansions associated with the SLAPT are more straightforward, as is apparent on comparing Eqns. (17–19) with their counterparts in [7, Appendix C]. With the SLAPT, there is no need to derive special expansions for the small angle case as is also done for the RAPT in [7, Appendix C]. Moreover, during numerical optimization of SLAPT parameters there is no need to perform co-ordinate conversion nor to include a barrier term in the objective function; as discussed in [7, §5.3], both of these are required for numerical optimization of the RAPT.

4. SPEECH RECOGNITION EXPERIMENTS

In this section we summarize the speech recognition experiments experiments undertaken to compare the effectiveness of RAPT- and SLAPT-based adaptation at reducing the word error rates of large vocabulary continous speech recognition (LVCSR) systems. These experiments were conducted using training and test material extracted from the *Switchboard Corpus*, a collection of approximately 2,500 conversations conducted over standard US telephone lines between two people 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 hours 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 used. This subset, dubbed *MsTrain*,[1] is composed of nearly 800 complete conversations spoken by 409 speakers, and totals 50.0 hrs. of speech. The test set used in all experiments 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 [3] along with first and second order difference coefficients derived from these.[2] 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 either a per-utterance or per-conversation side basis, as required by the particular experiment.

[1] The MsTrain set was provided by Dr. Joseph Picone and his students at Mississippi State University.

[2] The authors wish to thank Dr. Steve Young of Cambridge University for providing the implementation of PLP cepstral extraction used in these experiments.

Feature Normalization	% Word Error Rate	
	Full-Matrix MLLR	
	No	Yes
None	40.6	36.3
RAPT-1	38.8	34.8
RAPT-5	39.4	35.0
SLAPT-1	38.8	34.7
SLAPT-5	39.6	35.3

Table 1. Word error rates for systems trained with original and APT-normalized features, both with and without two-regression class MLLR/SAT.

All HMM training and test was conducted using HTK, the Hidden Markov Model Toolkit [10] as augmented by the Homewood Extensions. 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 generate a total of 6,712 state clusters in the final HMM. All word-error rates tabulated below were obtained by rescoring a set of trigram lattices with a modified version of the HTK decoding tool. The vocabulary used in generating and rescoring the lattices contained approximately 40,000 words.

Speaker Normalization

The first set of experiments was intended to establish the capacity of speaker-dependent (SD) normalization of cepstral features based on the APT to reduce the word error rate of a large vocabulary conversational speech recognition (LVCSR) system. In all cases reported below, feature normalization, when used, was applied to both test and training features to provide a matched condition. The normalization parameters for each speaker in the test and training sets were estimated with a simple GMM using the procedure delineated in [7, §6.1]. Single-pass retraining [10] was used to move the multiple-mixture HMM trained with the original cepstral features to the appropriately normalized set of features.

Feature normalization was tested in combination with MLLR. As before, MLLR, when used, was applied to both test and training; the basic SAT procedure [7, §4.4] was used for the latter. Adaptation of cepstral means was conducted with two fixed regression classes. In performing unsupervised estmation of MLLR parameters, an initial set of errorful transcripts obtained by decoding with the unadapted baseline system was used to perform the necessary forward-backward passes. This baseline system achieved a WER of 40.6%.

Table 1 reports results obtained with systems trained on the 50 hr. MsTrain set. From these results it is apparent that feature normalization with the one-parameter RAPT (i.e., the BLT) provides a WER reduction of approximately 1.5% absolute, and that this reduction is additive with that achieved using MLLR/SAT. Using APT-based feature normalization together with MLLR adaptation provided a total WER reduction of 5.8% beginning with an uncompensated system that achieved 40.6% WER. Also apparent is that normalization with the RAPT-5 transform provides some error rate reduction with respect to the un-normalized baseline, but that this reduction is not so large as that achieved using the

Enrollment Set	% Word Error Rate				
	RT-1	RT-9	ST-1	ST-9	MLLR
Baseline	41.5				
2.5 min.	38.5	37.3	38.4	37.4	37.1
60 sec.	38.3	37.4	38.2	37.5	37.5
30 sec.	38.5	37.6	38.3	37.7	37.9
10 sec.	38.7	37.8	38.6	38.0	40.1
5 sec.	38.8	37.9	38.6	38.2	45.5

Table 2. Results of rapid adaptation experiments with unsupervised enrollment data. RT (resp., ST) denotes the rational (resp., sine-log) all-pass transform.

simpler one-parameter transform. This result is somewhat counterintuitive: It may indicate that a simple GMM is not sufficient to estimate detailed, multi-parameter transforms.

Also reported in Table 1 are the results of a set of experiments undertaken to determine the WER reduction achievable with SLAPT-based normalization. These experiments used either a one- or five-parameter transform, once more in combination with MLLR/SAT. From these results we see that the RAPT- and SLAPT-based normalization schemes provide nearly identical WER reductions, whether or not MLLR/SAT is used in addition to feature normalization. As with the RAPT, one free parameter provides a more effective feature normalization than five free parameters.

Rapid Speaker Adaptation

We also tested the capability of the APT to reduce the error rate of an LVCSR when used for speaker adaptation. The results of a set of experiments conducted to compare full-matrix MLLR and APT-based adaptation on a task with limited unsupervised enrollment data are given in Table 2; in keeping with popular usage, we refer to this scenario as *rapid adaptation*. For these experiments, one global transformation was used for each speaker and CMS was applied on a per utterance basis. All systems were trained on the MsTrain set [7, §6.2]. The errorful transcripts used for unsupervised parameter, be it MLLR or APT, were obtained with the unadapted baseline system, which achieved a WER of 41.5%. As is apparent from the table, when 2.5 minutes of data were used during the unsupervised estimation of transformation parameters, the performance of MLLR and the nine-parameter APT systems were nearly identical. In this instance, the use of more free parameters in the all-pass transform resulted in further reductions in error rate. Also noteworthy is that as the amount of adaptation data was reduced, the performance of the MLLR system quickly deteriorated, suffering a catastrophic degradation at 10.0 sec. and less. The APT-based systems, on the other hand, experienced only marginal performance degradations, providing a reduction in WER of approximately 3.5% absolute with only 5.0 sec. of enrollment data. This difference in characteristics is surely due to the sparse parameterization of the APT.

5. CONCLUSIONS

In this work we have introduced the sine-log all-pass transform (SLAPT), a replacement for the rational all-pass transform (RAPT) considered in prior work. In a set of unsupervised speaker compensation experiments conducted on speech material from the Switchboad Corpus, both transforms were found to give very comparable, if not indentical, reductions in word error rate (WER). In a set of speaker normalization experiments, the gain from APT-based normalization was found to be additive with that provided by conventional maximum likelihood linear regression (MLLR). The combination of APT normalization with MLLR adaptation provided a error rate reduction of 5.8% absolute compared to an uncompensated baseline system which achieved 40.6% WER. In a set of unsupervised speaker adaptation experiments conducted on the Switchboard Corpus, MLLR- and APT-based systems were found to give nearly identical reductions in WER when an entire conversation side was used for speaker enrollment. As the amount of adaptation data was reduced, however, the performance of the MLLR system quickly deteriorated, suffering a catastrophic degradation at 10.0 sec. and less. The APT-based systems, on the other hand, experienced only marginal performance degradations, providing a reduction in WER of approximately 3.5% absolute with only 5.0 sec. of enrollment data from a baseline of 41.5%. This difference in characteristics is surely due to the sparse parameterization of the APT.

The Homewood Extensions (THE) are a set of C++ classes implementing the speaker adaptation and training algorithms discussed in this work; THE is publicly available for all non-commercial use at `isl.ira.uka.de/~jmcd`. THE has been ported to the Janus Speech Recognition Toolkit (JRTK), although a complete set of experimental results obtained with JRTK was not available at the time of publication.

6. REFERENCES

[1] A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 1990.

[2] Andreas Andreou, Theresa Kamm, and Jordan Cohen. Experiments in vocal tract normalization. In *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[3] Hynek Hermansky. Perceptual linear predictive (plp) analysis of speech. *J. Acoust. Soc. Am.*, 87(4):1738–1752, April 1990.

[4] 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.

[5] J. McDonough and W. Byrne. Speaker adaptation with all-pass transforms. In *Proc. ICASSP*, 1998.

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

[7] John W. McDonough. *Speaker Compensation with All-Pass Transforms*. PhD thesis, The Johns Hopkins University, Baltimore, MD, 2000.

[8] John W. McDonough and William Byrne. Single-pass adapted training with all-pass transforms. In *Proc. Eurospeech*, 1999.

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

[10] S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK Book*. Entropic Software, Cambridge, 1999.