SPEAKER NORMALIZATION THROUGH FORMANT-BASED WARPING OF THE FREQUENCY SCALE

Evandro B. Gouvêa and Richard M. Stern

Department of Electrical and Computer Engineering
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213, USA
Tel. +1 412 268 7116, FAX: +1 412 268 3890, Email: egouvea@cs.cmu.edu

ABSTRACT

Speaker-dependent automatic speech recognition systems are known to outperform speaker-independent systems when enough training data are available to model acoustical variability among speakers. Speaker normalization techniques modify the spectral representation of incoming speech waveforms in an attempt to reduce variability between speakers. Recent successful speaker normalization algorithms have incorporated a speaker-specific frequency warping to the initial signal processing stages. These algorithms, however, do not make extensive use of acoustic features contained in the incoming speech.

In this paper we study the possible benefits of the use of acoustic features in speaker normalization algorithms using frequency warping. We study the extent to which the use of such features, including specifically the use of formant frequencies, can improve recognition accuracy and reduce computational complexity for speaker normalization. We examine the characteristics and limitations of several types of feature sets and warping functions as we compare their performance relative to existing algorithms.

1. INTRODUCTION

In recent years nonlinear frequency normalization has become a popular approach to reducing the effects of systematic variations of vocal tract anatomy on speech recognition accuracy. Typically frequency normalization involves the use of a “warping function” which is intended to characterize a mapping of the average spectra of two speakers (or between a given speaker and a “standard” speaker). Two key issues that distinguish the various frequency warping are the shape of the warping functions used and the method by which they are selected. For example, linear warping functions have been chosen by many research groups, in part because of their simplicity. Other common choices are curves based on speech perception studies, such as the bilinear transform or transforms based on the mel scale, along with curves based on speech production models. The selection of warping function is sometimes accomplished by choosing from a set of candidate functions in a fashion that maximizes the likelihood of the observations, and sometimes based directly on speaker-specific speech parameters.

In a relatively early study, Acero blindly estimated the optimal frequency-distortion parameter for the bilinear transform used to accomplish frequency warping for LPC-derived cepstra [1, 2]. This technique produced a 12 percent decrease in the relative error rate on the CMU speaker-independent alphanumeric census task.

Much of the current activity in speaker normalization was motivated by the results of Cohen *et al*. [3], who used a set of linear frequency warping functions. They chose specific warping functions by training HMMs on half the data using the current warping function and applying the decoder to the other half. The warping function was chosen by iteratively searching for the shape that maximized the likelihood of the decoder output, interchanging the roles of the two subsets of data during each iteration.

Wegmann *et al*. [9] described a less computationally-intensive approach. They used a Gaussian mixture model to represent the features of the standard speaker, and a piecewise-linear warping function was chosen to maximize the likelihood of these features (rather than the likelihood of the decoder output).

Lee and Rose [7] proposed a technique that combines some of the better features of the techniques of Cohen *et al.* and Wegmann *et al.* They choose the best warping function for speakers in the training set using HMMs and, after an iterative process, estimate a Gaussian mixture model which will be used to select the warping function for speakers in the test set.

A completely different approach was adopted by Eide and Gish [5]. They estimated a parameter k_s based on the third formant for each speaker and used as the warping function

$$F = k_s^{\frac{3f}{8000}} f \qquad (1)$$

Zhan and Westphal [10] report on extensions of the algorithms proposed by Lee and Rose and by Eide and Gish. They report results using a piecewise-linear warping function, as well as a function similar to Eq. (1). For both sets of warping functions, that is, for linear and nonlinear warping functions, they find the appropriate constants using two methods: they compute the constants directly from each of the three first formants, applied separately (in a fashion similar to Eide and Gish), or they search a range of values and choose the best one by maximizing of likelihoods (similar to Lee and Rose).

None of the above approaches makes extensive use of acoustic features. In this paper, we explore the use of frequency warping techniques that are based more directly on the observed values of ensembles of acoustic features. Specifically, we examine the performance of frequency warping functions based on the first three formant frequencies. In general, we compile histograms of formant frequencies from speech from a particular speaker in the test set, and we use features such as the median, maximum, or minimum of each formant frequency to derive the warping function.

In Section 2, we outline the basic postulates of this approach. In Section 3, we describe the results of several pilot experiments that demonstrate the algorithm’s effectiveness and that compare its performance to other normalization procedures in terms of recognition accuracy and computational requirements. In Section 4 we describe several generalizations of the algorithm that use different warping function shapes or different feature sets.

2. DESCRIPTION OF THE ALGORITHM

The two most important issues to be resolved in nonlinear frequency warping concern the general shape of the warping function and how its specific form is obtained. As noted above, we select the warping function based on points extracted from acoustic features, specifically distributions of formant frequencies, on a speaker-by-speaker basis. The specific warping function is obtained by comparing the values of these features obtained for a specific speaker with the corresponding feature values averaged over all speakers in the training set.

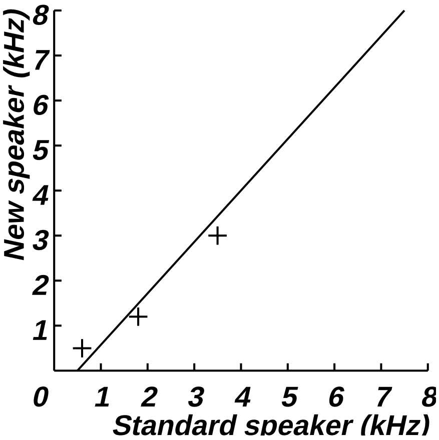

Figure 1. Example of warping function based on linear fit to points extracted from formants.

Figure 1 demonstrates a warping function obtained in this fashion from three hypothetical formant-based features. The ordinate of each of the three points is a statistical parameter computed from the histograms of the three formants from a particular speaker. The abscissa represents the same statistical feature values from the "standard speaker", which is obtained by computing the means of the features used over all speakers in the training set. A warping function is fitted to these three points, which in this case is assumed to be linear. Since we use a sampling frequency of 16 kHz in our work, the maximum frequency shown, 8 kHz, is the Nyquist frequency.

3. PERFORMANCE OF FORMANT-BASED FREQUENCY NORMALIZATION

In this section we describe the results of several experiments which demonstrate the performance of formant-based frequency normalization and compare it to other approaches. We use the medians of the three formant frequencies as features, and a linear warping function that is not constrained to include the origin for the experiments described in this section. In Section 4 below, we describe the results of some pilot experiments in which we consider a more general set of warping functions and features.

3.1. Distribution of Slopes of Linear Warping Functions According to Gender

Since females usually have vocal tracts with smaller dimension than males, a female's formant frequencies are usually higher than a male's. We would expect that a reasonable warping function would cause a female's formants to become compressed in the process, while a male's formants would be expanded, so that the warped frequencies are closer together. Compression can be achieved by a warping function that is a straight line with a slope greater than one, and expansion, with a slope that is less than one.

To confirm this expectation we calculated a series of speaker-specific linear warping functions as described in the previous section and plotted histograms of their slopes, separated by gender. Figure 2 shows these results. We note that the clusters separate well by genders, and that the dependence of slope on gender is as expected.

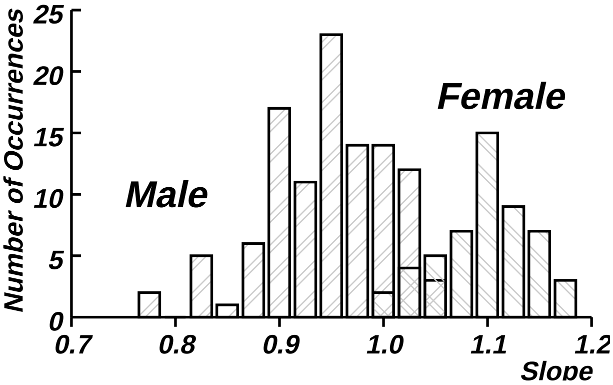

Figure 2. Distributions of slopes of linear warping functions based on best fit through median formant frequencies according to gender.

3.2. Effect of Normalization on Formant Clusters

We also examined the extent to which the frequency normalization process reduces the variability of formant frequencies across all speakers. Figure 3 represents a replotting of the data collected by Peterson and Barney [8], who estimated formant frequencies from vowels uttered by approximately sixty speakers, each speaker repeating each of ten vowels twice. We applied our frequency normalization algorithm to the Peterson-Barney data, and plotted points in the formants planes, before normalization and after normalization. For clarity's sake, we plot just three of the ten vowels considered in Figure 3, and only data from 31 speakers. The three vowels plotted establish the vertices of the "vowel triangle" proposed by Peterson and Barney.

The Fisher ratio, which is the ratio between the variance of the means divided by the means of the variances across several different clusters, is a commonly-used measure of cluster separation relative to intrinsic cluster variability. The Fisher ratio for the clusters of data replotted in Figure 3 is 12 before normalization and increases to 29 after normalization. Hence, from a pattern recognition perspective, the transformation is making the classes more compact and farther apart, which would provide greater classification accuracy.

3.3. Effect of Frequency Normalization on Recognition Accuracy

To verify that frequency normalization is effective, we ran a series of recognition experiments using the linear warping function and the median values of the first three formants as features. The results were obtained using the speaker independent portion of the Resource Management database (RM1). The training set has 120 speakers, including 85 males and 35 females. The test set has 40 speakers, 23 males and 17 females. Training and testing were performed using SPHINX-III, which is a continuous HMM system, with output probability distributions consisting of 2-component Gaussian mixtures.

Table 1 summarizes our results, comparing recognition accuracy using formant-based frequency normalization with a local implementations of the algorithm proposed by Wegmann *et al.* [9] that was used by CMU in the 1995 ARPA Hub 3 continuous speech recognition evaluation. Our implementation of the algorithm of Wegmann *et al.* is described in [6], and it differs in the shape of the warping function from the original implementation

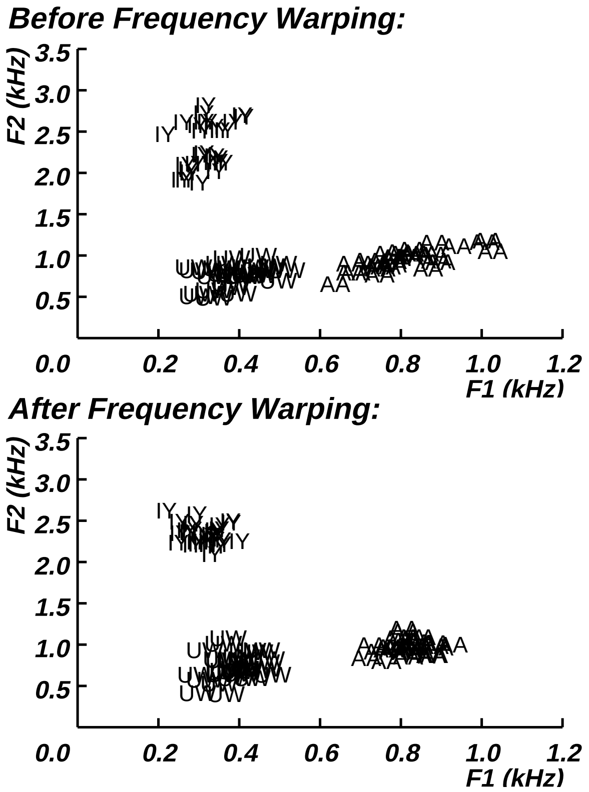

Figure 3. *Formants of the phonemes /AA/, /IY/, and /UW/ before normalization (upper panel) and after normalization (lower panel). Data collected by Peterson and Barney [8].*

proposed by Wegmann *et al.* [9]. As can be seen in Table 1, the use of formant-based frequency warping produced a greater reduction in word error rate for this dataset than our implementation of the algorithm proposed by Wegmann *et al*.

Method	Error Rate	Improvement
Baseline (No Normalization)	6.6%	–
Frequency Warping Based on Wegmann *et al.*	6.3%	4.5%
Formant-Based Frequency Warping	5.8%	12%

Table 1. *Comparison of word error rate using formant-based frequency warping with own implementation of the algorithm of Wegmann* et al.

3.4. Computational complexity

In the recognition results presented in the previous section, we estimated medians of formants for each speaker using all 40 sentences in the data. Figure 4 describes how recognition accuracy is affected by the number of sentences used to estimate the median formants. As can be seen, 5 sentences appear to be adequate.

Table 2 describes the computational time required to enroll a new speaker to the system using a Hewlett-Packard HP-720 workstation. 15 sentences from each speaker were used for both the formant-based normalization and our implementation of the method of Wegmann *et al.* It can be seen that the formant-based frequency warping is faster as well as more accurate.

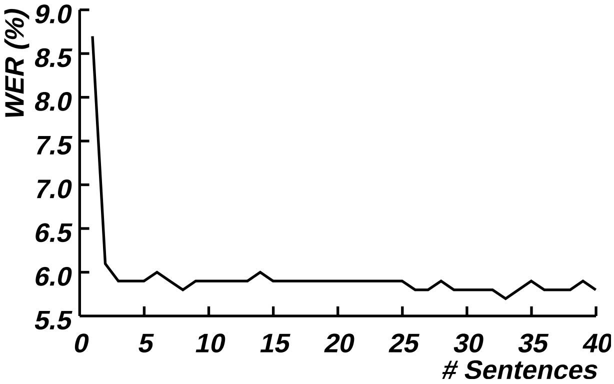

Figure 4. *Word error rate as a function of the number of sentences used to estimate medians of formants for speakers in the test set.*

Method	Error Rate	Enrollment Time
Formant-Based Frequency Warping	5.9%	175 s
Frequency Warping Based on Wegmann *et al.*	6.3%	251 s

Table 2. *Comparison of enrollment times using formant-based frequency warping and our implementation of the method of Wegmann et al. 15 enrollment sentences were used.*

3.5. Robustness to Additive Noise

Additive noise can affect the performance of formant-based frequency warping, both by degrading the quality of the feature estimates and by degrading the ability to estimate formant frequencies. We observed the effect of noise on performance by artificially adding white Gaussian noise at different signal-to-noise ratios (SNRs) to speech from the speaker-independent RM1 sentences, and then estimating formants, and recognizing as before.

SNR	5 dB	10 dB	20 dB	40 dB
No Warping	89.1	71.0	23.8	6.6
Formant-Based Warping	89.0	69.1	20.3	5.8

Table 3. *Word error rates for speech with additive noise at several SNRs.*

Table 3 describes word error rate for several SNRs, before and after normalization. Formant-based frequency warping consistently provides better word error rates at all SNRs considered, with greatest percentage improvement at high SNRs.

4. GENERALIZATIONS OF THE ALGORITHM

The results presented in the previous section show that formant-based frequency warping using a linear warping function and the medians of the first three formants as features can achieve the same or better recognition accuracy at lower computational cost than the similar algorithms to which it was compared. In this section we briefly summarize some results obtained using a more general class of warping function shapes and a more general class of features.

4.1. Warping Functions from Normal Deviate Transformations

A wider variety of potential warping functions could be obtained by considering the shapes of typical Receiver Operating Characteristics (ROC or “isosensitivity” curves), which are plots of detection versus false alarm probabilities (*e.g.* [4]). If the underlying statistics of the received signal in a communications system are Gaussian, these ROC curves become linear when the probabilities are replaced by their corresponding “normal deviates”. The normal deviate Z of a probability P is defined implicitly as

$$P = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x \qquad (2)$$

We obtained a more general set of formant-based warping functions by first normalizing formant frequencies by dividing them by the Nyquist frequency. We then applied the normal deviate transformation (Eq. 2) to the normalized frequencies (which now had values between 0 and 1), determined the best-fit straight line in the Z-domain, and computed the inverse normal deviate transformation (*i.e.* the Gaussian error function) of the values of Z defined by the resulting curve.

The use of warping functions derived from best-fit straight lines in the normal deviate domain reduced the word error rate on the RM1 data from 5.8% to 5.6%, which increased the percentage improvement relative to the baseline from 12% to 15%. Some individual speakers obtained lower error rates using warping functions that were linear in the original frequency domain, while others obtained better results with warping functions obtained using the normal deviate transform. While selection of the better-performing warping function for each speaker would have reduced the error rate to 5.3% (a 20% reduction relative to the baseline), we could not identify a way to do so blindly.

Figure 5 illustrates warping functions obtained using original linear and the normal deviate approach for a Speaker CEG0 in the RM1 database. As can be seen, the two curves are similar for lower frequencies but they diverge at higher frequencies.

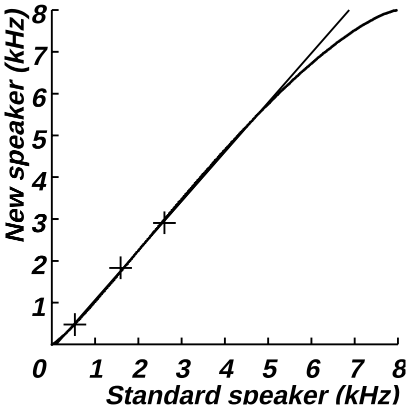

***Figure 5.** Comparison of warping functions obtained using the linear and normal-deviate approaches for a speaker in the RM1 database*

4.2. Use of Additional Features

We also explored the use of other statistical parameters extracted from formant frequency histograms. Because the extremal formant frequencies are believed by some to be instrumental in defining the shape of the “vowel triangle”, we evaluated the performance of warping functions obtained from features consisting of the 5th and 95th percentiles of each of the first three formant frequencies (rather than the median or 50th percentile). We considered both the linear and the normal-deviate-based warping functions. Nevertheless, we obtained error rates that were worse than the baseline in each case considered.

5. CONCLUSIONS

This paper presents a new approach to speaker normalization that takes advantage of the medians of the first three formant frequencies to find a suitable nonlinear frequency warping function. Our pilot results indicate that the proposed method provides error rates that are slightly lower than other existing approaches, and with reduced computational requirements.

ACKNOWLEDGMENTS

The authors thank Bhiksha Raj and Pedro Moreno for useful suggestions and the rest of the speech group for the encouragement. Evandro Gouvêa has been supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil. This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

REFERENCES

[1] Acero, A., *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, MA, 1993.

[2] Acero, A. and Stern, R. M., “Robust Speech Recognition by Normalization of the Acoustic Space” *Proc. ICASSP-91*, pp. 893-896, May 1991.

[3] Cohen, J., Kamm, T., Andreou, A., “An Experiment In Systematic Speaker Variability”, Final Day Review., DoD Speech Workshop on Robust Speech Recognition, Baltimore, MD, August 1994.

[4] Egan, J. P., *Signal Detection Theory and ROC-Analysis*, Academic Press, New York, 1975.

[5] Eide, E., Gish, H., “A Parametric Approach To Vocal Tract Length Normalization”, *Proc. ICASSP-96*, Vol. 1, pp. 346-348, May 1996.

[6] Gouvêa, E. B., Moreno, P. J., Raj, B., Stern, R. M., “Adaptation and Compensation: Approaches to Microphone and Speaker Independence in Automatic Speech Recognition”, *Proc. DARPA Speech Recognition Workshop*, pp. 87-92, February 1996.

[7] Lee, L., Rose, R. C., “Speaker Normalization Using Efficient Frequency Warping Procedures”, Proc. ICASSP-96, Vol. 1, pp. 353-356, May 1996.

[8] Peterson, G. E., Barney, H. L., “Control Methods Used In A Study Of Vowels”, *J. Acoust. Soc. Am.*, Vol. 24, pp. 175-184, 1952.

[9] Wegmann, S., McAllaster, D., Orloff, J., Peskins, B., “Speaker Normalization On Conversational Telephone Speech”, *Proc. ICASSP-96*, Vol. 1, pp. 339-341, May 1996.

[10] Zhan, P., Westphal, M., “Speaker Normalization Based on Frequency Warping”, *Proc. ICASSP-97*, Vol. 2, pp. 1039-1041, April 1997.