A PARAMETRIC APPROACH TO VOCAL TRACT LENGTH NORMALIZATION

Ellen Eide and Herbert Gish
BBN Systems and Technologies
70 Fawcett St. 15/1c
Cambridge, MA 02138 USA

ABSTRACT

Differences in vocal tract size among individual speakers contribute to the variability of speech waveforms. The first-order effect of a difference in vocal tract length is a scaling of the frequency axis; a female speaker, for example, exhibits formants roughly 20% higher than the formants of from a male speaker, with the differences most severe in open vocal tract configurations.

In this paper we describe a parametric method of normalization which counteracts the effect of varied vocal tract length. The method is shown to be effective across a wide range of recognition systems and paradigms, but is particularly helpful in the case of a small amount of training data.

INTRODUCTION

One way to compensate for differences in vocal tract length has been described by [1], who performed an exhaustive search for the best linear scaling of the frequency axis, $f' = k_s f$, for each training speaker s. Models were iteratively developed by choosing the scale factor k_s which maximized the likelihood of the training speaker's acoustics in the current model.

For each test speaker, decoding was done at each of twenty choices for the scale factor. The recognition associated with the maximum likelihood factor was chosen as the system output.

The results showed that appropriate scaling of the frequency axis for each speaker can significantly reduce the number of errors produced by a speech recognition system. However, the exhaustive, iterative approach outlined above is too costly for reasonable amounts of training data; decoding each test speaker twenty times is similarly limited.

In our paper, we will present a different approach to vocal tract length normalization. We derive an *estimate* of the appropriate scale factor for each speaker based on formant positions, thereby enabling inexpensive computation of arbitrary precision. We show the results of the procedure under a linear as well as a non-linear warping function.

METHOD

We introduce into our analysis program a function g which transforms the frequency axis according to $f' = g(k_s, f)$ where k_s is a scalar which compensates for the vocal tract length of speaker s. We first motivate our choices for g based on simple vocal tract models and then describe how k_s is estimated.

Consider the simple vocal tract models shown in figures 1 and 2. The simplest, the uniform tube, is appropriate for relatively open vowels such as /AA/. For this model, formant frequencies occur at odd multiples of $1/L$; a change in the length of the vocal tract from L to kL results in a scaling of the frequency axis by a factor $1/k$, which is consistent with a linear warping. Following the uniform tube model, then, the frequency axis is warped according to $f' = k_s f$. However, a uniform tube is not always the best model of the vocal tract. For example, the model depicted in figure 2, known as a Helmholtz resonator, is a good approximation of the vocal tract configuration for the first formant of the close front vowel /IY/. The resonance generated by such a system, F_1, is proportional to $\sqrt{V/AL}$, where V is the volume of the back cavity, A is the cross-sectional area, and L is the length of the narrow tube. Fant [2] suggests that a possible way to model a change in overall vocal tract size by a factor k is to adjust all dimensions of the model by k except the length L of the narrow tube, which remains fixed. The resulting shift of the first formant frequency is $1/\sqrt{k}$ making the first formant region of these vowels less sensitive to a change in vocal tract length. Furthermore, the higher frequency regions for these vowels show sensitivity to k comparable to that

0-7803-3192-3/96 $5.00

of open vowels.

Thus, the scaling of the frequency axis imposed by a change in vocal tract length is dependent on the configuration of the vocal tract or, equivalently, on the phoneme being produced, with the formant positions of open vowels most affected by changes in tract length.

In order to compromise among the various phoneme-dependent effects of vocal tract length in a context-independent normalization scheme, we have investigated a warping of the frequency axis according to $f' = k_s{}^{3f/8000} f$.

This non-linear warping, which allows more stretching at high frequencies than at low, provides slightly better performance than a linear warp.

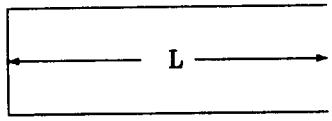

Figure 1. Uniform tube model.

Having decided on a form for g, we derive a single normalization factor for each training and testing speaker based on the median position of the third formant in his speech. Formants are estimated using the commercially-available Waves+ package [3]. This software solves for roots of the LPC equation at each frame, chooses the desired number of formants from the set of roots, and then smooths the estimated formant tracks through an utterance using dynamic programming. For male speakers we search for four formants in the range 300-3300 Hz, while for females we look for three formants in this range.

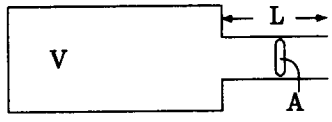

Figure 2. Helmholtz resonator.

Having estimated a formant configuration for each frame of each waveform from a given speaker s, we compute his normalization factor, k_s. This factor is calculated as the median of the speaker's third formant over a subset of frames satisfying the criteria itemized below, divided by the median of F_3 under the same constraints taken across the set of training speakers. The criteria used to include frames in the median process were:

- $p_v > 0.8$
- $F_1 > 400Hz$
- $2kHz < F_3 < 3kHz$

where p_v is the probability of voicing at a given frame as specified by Waves+.

To reduce the computation, rather than use all of the speech from each speaker to estimate the median position of the third formant, we have used only the first 30 utterances from each speaker, where the average utterance lasts 2.6 seconds.

EXPERIMENTS AND RESULTS

Our normalization method has performed well over a variety of experimental paradigms on the Switchboard corpus. The method shows the largest gains over baseline at small amounts of data, but significantly increases performance at large amounts of training as well.

Paradigm	Error
[1.0] Baseline	61.9
[1.1] Linear forced	59.4
[1.2] Linear optional	57.3
[1.3] Non-linear forced	58.0
[1.4] Non-linear optional	57.1

Table 1. Effect of normalization on the CAIP test set. 5 hours of training. 5k closed vocabulary.

In conditions [1.1] and [1.2] we perform the mapping $f' = k_s f$. Condition [1.1] forces the use of the estimated normalization factor for each test speaker. We have noticed that the recognition results are very sensitive to k_s; a small change in this value in the wrong direction can cause a large increase in the error rate for that speaker. To diminish this effect, we allow in testing an optional stretch on a sentence-by-sentence basis. Each sentence is decoded twice, once with the estimated normalization factor included in the analysis, and once with k_s set to 1. We select as our recognition hypothesis the transcription associated with the higher of the two likelihoods. Results of enabling this type of back-off are shown in condition [1.2]. Note that the backing-off is allowed only in decoding; all training speakers are normalized according to their estimated scale factor.

In conditions [1.3] and [1.4] we perform the mapping $f' = k_s^{(3f/8000)} f$. Condition [1.3] uses the estimated normalization factor while condition [1.4] allows the option of backing off to a factor of 1.0 for k_s. We see that in this test, after allowing back-off, there is little difference between the two choices for g. Nonetheless,

Paradigm	Test Unnorm	Test Non-linear Forced	Test Non-linear Optional
Train Unnorm	60.2	58.6	58.1
Train Non-linear forced	57.3	56.0	55.2

Table 2. The effect of normalizing only the training or test set. 10k words + 2k compounds open vocabulary. 5 hours of training data.

we assume the non-linear form for the remainder of the paper.

We have also looked at the results when only the training set or the test set is normalized. These results are collected in table 2 for the case of 5 hours of training data. We note that the gains from normalizing the training and the test are nearly additive. In this small-training problem, most of the gain arose from normalizing the training data which enabled more efficient use of the limited data.

Paradigm	Hours of Training		
	5	32	63
[3.0] Baseline	60.2	51.9	51.8
[3.1] Non-linear, forced	56.0	49.6	49.3
[3.2] Non-linear: optional	55.2	48.8	48.8

Table 3. Effect of normalization on the CAIP test set as a function of training data. 10k open vocabulary plus 2k compound words.

Next, we looked at performance as a function of the amount of training data used, as shown in table 3. Note that the normalization procedure is especially effective in the case of small amounts of training. Furthermore, under normalization the results saturate earlier than in the unnormalized case. We could exploit this fact two ways. We could use less data for training, thereby requiring less training time. Alternatively, we could use all of the data but increase the complexity of our models and achieve potentially better recognition performance. This option is as yet unexplored.

Finally, we looked at the effect of simulating training data by warping each training speaker by multiple factors and pooling the results into an enlarged training corpus, as shown in table 4. Condition [4.0] is the baseline case of a single normalized training token for each original training observation with optional scaling in test. Condition [4.1] reflects the pooling of the non-linearly warped data of condition [4.0] together

Paradigm	Error
[4.0] Non-linear optional	55.4
[4.1] Non-linear forced pooled	56.7
[4.2] Non-linear optional pooled	55.8

Table 4. The effect of simulating training data, using k_s and $(k_s+1.0)/2$ as normalization factors for speaker s. An open vocabulary of 10k words + 2k compounds was used.

with a normalization in which the normalization factor was taken as $(k_s + 1.0)/2$, with the test data warped according to k_s. Condition [4.2] indicates the performance using the pooled-data model and allowing an optional warping of the test set. As indicated, generating new training samples does not reduce the error rate over just using a single, properly normalized data set for training. This is easily explained by the fact that the pooled models are smeared relative to the single-warping case, with no new contextual information provided as would be the case for increasing the amount of actual Switchboard data used for training.

SUMMARY

In this paper we have described a method of warping the frequency axis to compensate for the vocal tract length of each training and testing speaker. The procedure is tractable for large amounts of data in that it estimates the appropriate scale factor for each speaker based on formant positions. We have shown results on the Switchboard corpus for two choices for the warping function, and we have shown the effect of the procedure as a function of the amount of training data available. We have found that although effective at all levels of training data, it is particularly so for small problems. Finally, we found that simulating training data by multiple scalings of the frequency axis did not help recognition performance.

REFERENCES

[1] Kamm,T., Andreou, A., and Cohen, J. *Vocal Tract Normalization in Speech Recognition: Compensating for Systematic Speaker Variability.* Proceedings of the Fifteenth Annual Speech Research Symposium. Baltimore, MD. June, 1995.

[2] Fant, G. *Speech Sounds and Features.* Cambridge, MA: The MIT Press. 1973.

[3] Waves+. Version 5.0. Entropics Research Laboratory. 1993.