INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 666-2900 • FAX (510) 666-2956

Pitch-based Vocal Tract Length Normalization

Arlo Faria *

TR-03-001

November 2003

Abstract

This paper investigates the correlation between fundamental frequency and resonant frequencies in speech, exploiting this relation for vocal tract length normalization (VTLN). By observing a speaker's average pitch, it is possible to estimate the appropriate frequency warping factor which will transform a spectral representation into one with less variation of the formants. I use a function of pitch that maps to a corresponding frequency warping factor. An exploration of speaker and vowel characteristics in the TIMIT speech corpus is used to optimize the parameters of this function. The approach presented here is a potentially simpler alternative to existing VTLN algorithms which derive the warping factor by other means. Recognizer results indicate that the pitch-based approach compares favorably against other methods; furthermore, performance could be further improved by using a warping function that is not strictly linear.

* The author is a student in the Department of Electrical Engineering and Computer Science at the University of California, Berkeley.

This research was supported by SUPERB-IT and was performed under the supervision of Prof. Nelson Morgan and David Gelbart at the International Computer Science Institute in Berkeley, California.

Index Terms

Vocal tract (length) normalization, speaker adaptation, speaker normalization.

I. Introduction

AUTOMATIC speech recognition systems generally strive to be *robust* – capable of performing under adverse conditions such as noise and unaffected by the considerable diversity of speech produced by different speakers in a population. Toward this latter goal, vocal tract length normalization (VTLN) is an attempt to improve recognition perfomance by compensating for some of these differences and transforming the system's input accordingly.

The sizes of speakers' vocal tracts is a physiological factor which is directly related to the vocal tract's resonant frequencies, phonetic attributes known as formants. Phone discrimination, particularly of vowels, relies on observations of the formants, especially the lowest two (F_1 and F_2). Small variations in these frequencies are not always salient to human auditory perception but can affect the pattern classification of an automatic speech recognition system. To address this complication, VTLN is utilized to predict the vocal tract length of a speaker and appropriately rescale the frequency axis of the spectral representation.

This paper reviews some existing VTLN schemes and introduces a novel pitch-based approach which performs comparatively well. Contemporary approaches are either based on speaker characteristics that are difficult to measure accurately or involve a complex probabilistic procedure that is not straightforward to implement. The proposed VTLN approach is simple and effective, and could be an easily integrated benefit to a speech recognition system.

The following sections of this paper briefly discuss the major approaches to VTLN before presenting the pitch-based approach. I then explain the determination of warp factors as a function of pitch, along with several ways to optimize the parameters of such a function. Additionally, frequency warping functions are examined, and I present an alternative method of realizing warping functions by directly processing the input speech signal prior to feature calculation.

Lastly, some results from recognizers trained with pitch-based VTLN demonstrate that the proposed approach is feasible; moreover, the pitch-based approach is capable of considerable gains. Experiments with the alternative warping function suggest that this promising performance could even be improved.

II. Approaches to VTLN

Approaches to vocal tract length normalization differ in the processes by which they calculate the warp factors that are used to transform the input to a speech recognizer. The existing approaches can be considered in two classes: the Maximum Likelihood (ML) approach choses the warp factor that makes a speaker's utterances most probable in a given model; other approaches are based on speaker-specific acoustic characteristics that are intrinsically related to the vocal tract length.

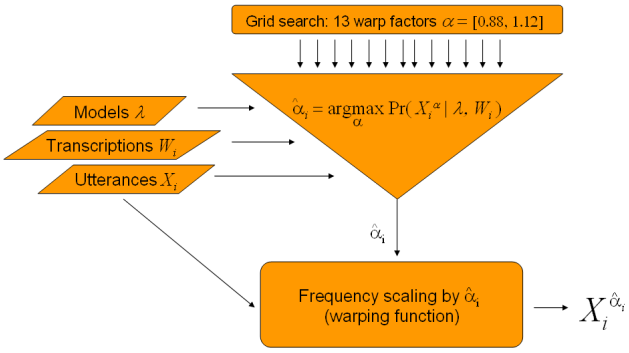

Fig. 1. The ML approach to VTLN

A. Maximum Likelihood warp factors

The Maximum Likelihood method [1] [2] [3] for determining warp factors involves maximizing the likelihoods of speakers' utterances based on a given model, usually a Hidden Markov Model. For a speaker i, given a set of models λ along with a set of utterance observation vectors X_i and their transcriptions W_i, the optimal warp factor $\hat{a}_i$ is defined:

$$\hat{a}_i = \underset{a}{\text{argmax}} \Pr\left(\mathrm{X}_i^a | \lambda, W_i\right) \tag{1}$$

where a is the factor that is applied to create the warped observations X_i^a. Because Equation 1 is difficult to solve analytically, the optimal $\hat{a}_i$ is found by a search over a discrete set of factors [1]. The process is illustrated in Figure 1.

The training procedure for ML VTLN is fairly intricate and usually consists of a *two-pass* process: half of the train set is used for training an HMM while $\hat{a}_i$ is found for the other half; then the sets are swapped. This is iterated until the estimated $\hat{a}$s are stabilized between iterations. The testing procedure involves decoding with every discrete value for a and warping with the $\hat{a}$ that maximizes the probability of the X_i^a given the normalized model λ_N.

The primary advantage of the ML approach is that it is designed to find a warp factor that is optimal, in a probabilistic sense. However, the process is computationally intensive, especially as the resolution of the warp factor search grid increases. It is worth noting, however, that much work has been done to streamline the ML approach [3], and that the argument against computational inefficiency becomes less severe as computing power grows with time. The greatest disadvantage of this approach – from an implementation standpoint – is its complexity and scalability.

Nonetheless, the ML approach works rather well and is regarded as a standard for VTLN.

B. Formant-based approaches to VTLN

Unlike the ML approach which has no acoustic motivation, there are a number of VTLN approaches which calculate the warp factors by observing speaker-specific

speech characteristics. Most commonly, the warp factors are based on the ratio of speakers' third formants, F_3. The reason for this is well-founded, given the dependency of resonant frequencies on vocal tract length according to the open-tube model. Presumably, the third formant should be influenced by vocal tract length in the same manner as the first two formants. The third formant is chosen because it is fairly constant, not changing as much as the more phonetically informative lower formants.

In [4], a warp factor based on F_3 is used to improve recognition on mismatched train and test sets. For example:

$$w = \frac{F_{3,children}}{F_{3,adults}} \tag{2}$$

For a more general warp factor, a speaker's average F_3 is compared to the average F3 of the entire speaker set [5]:

$$k_s = \frac{\overline{F_{3,speaker}}}{\overline{F_{3,all}}} \tag{3}$$

Warp factors based on F_3 seem to have a very strong relation to the vocal tract length – and thus the appropriate scaling of the first and second formants. In practice, however, this approach is limited by the difficulty of isolating the third formant in a spectrum. Algorithms for tracking formants are not always reliable, and a misestimated formant could result in a substantial magnitude of error.

III. Evidence of pitch-formant correlation

The motivation for the pitch-based approach also has a physiological basis; unlike the formant-based warp factors discussed previously, though, fundamental frequency (pitch) is not a characteristic of vocal tract length and does not directly affect the resonant frequencies. Rather, fundamental frequency represents the rate of vocal cord excitations. Thus pitch is more closely related to the physical size of a speaker's larynx.

Pitch-based VTLN exploits the assumption that speakers with large voiceboxes also have large vocal tracts; similarly, speakers with higher F_0 will have smaller larynxes along with shorter vocal tracts, and thus higher formant frequencies. Therefore, to predict frequency warping factors it is adequate to observe a speaker's pitch.

To investigate this effect, one can explore the vowel characteristics in a large group of speakers:

- For this survey, the TIMIT corpus was used. This is a phonetically balanced corpus of 630 speakers from eight dialect regions of the United States.
- The vowels considered (transcribed /iy/, /ae/, /aa/, /uw/) occurred in utterances `sa1` and `sa2` of the corpus, corresponding to the two common sentences that every speaker read. This assured that all phones were in a similar context.
- A pitch-tracking tool[1] was used to estimate a speaker's median F_0 over the duration of each utterance

[1] The pitch-tracking tool is called *get_f0*, in a software package distributed and copyrighted by Entropic Research Laboratory

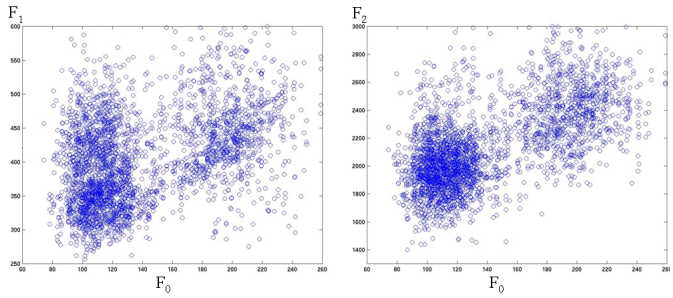

Fig. 2. Pitch versus formants F_1 and F_2 of the vowel /iy/

- A formant-tracking tool[2] was used to locate the first and second formants.

This experiment demonstrated a clear correlation between the pitch and formants as seen in Figure 2, where the fundamental frequency is plotted against the F_1 and F_2 of the high-front vowel /iy/[3].

IV. Determination of a pitch-based warp factor

A. *The warp factor k as a function of F_0*

Considering the correlation between pitch and formants examined above, it is possible to derive appropriate warp factors that will transform the formants according to the vocal tract size estimated from pitch. This indicates a warp factor k that is a function of F_0:

$$k = f(F_0)$$

Suppose that it were possible to average the frequencies of a certain formant for a particular speaker and characterize that as F. In applying VTLN, the warp factor k is the multiplicative constant that accounts for the inter-speaker frequency variations of the formants F and can be used to shift those formants to the normalized F′:

$$F' = kF \tag{4}$$

Given the correlation evident in Figure 2, it is possible to have a function y that represents the linear best fit, minimizing error in a least-squares sense. This represents the approximate locations of a particular vowel's formants, as a function of F_0:

$$F \approx y(F_0) = aF_0 + b \tag{5}$$

To determine the function for the pitch-based warp factor, first consider that the function y has a fixed point at which $y(F_0)$ will be equal to the desired F′, representing the mean frequency to which all other F are normalized. The value of F_0

[2] The tool is *formant*, from Entropic Research Laboratory

[3] The vowel /iy/ was chosen for inspection because its formants are spread widely and less susceptible to error in the formant-tracking tool. Also, /iy/ is the most common vowel in the TIMIT corpus.

corresponding to this point is μ, the "average" pitch for which the VTLN warp factor equals 1.

$$\mathrm{F}' = y(\mu) \tag{6}$$

Combining (4), (5), and (6) gives the desired k as a function of F_0, along with linear fit coefficients a and b for a particular formant, and the parameter μ:

$$k = f(\mathrm{F}_0) = \frac{y(\mu)}{y(\mathrm{F}_0)} = \frac{a\mu + b}{a\mathrm{F}_0 + b} \tag{7}$$

B. A linear approximation for k

To simplify the process of locating optimal parameters for the equation in (7), it is helpful to represent the function's linear approximation. This approximation should be done at the value μ to minimize the errors at the extremes of the fundamental frequency domain. Given the function f defined in (7):

$$k = f(\mathrm{F}_0) \approx \mathrm{L}_\mu(\mathrm{F}_0) = f(\mu) + f'(\mu)(\mathrm{F}_0 - \mu)$$

$$f'(\mathrm{F}_0) = \frac{-a(a\mu + b)}{(a\mathrm{F}_0 + b)^2}$$

$$\mathrm{L}_\mu(\mathrm{F}_0) = 1 - \frac{a}{a\mu + b}(\mathrm{F}_0 - \mu) \tag{8}$$

The function f and its linear approximation L_μ are compared in Figure 3, which uses the linear fit coefficients calculated for the second formant of /iy/:

$$\mathrm{F}_{2,/iy/} \approx a\mathrm{F}_0 + b = 4\mathrm{F}_0 + 1600 \tag{9}$$

C. Finding optimal parameters

Rewriting the slope of (8) as α,

$$k = f(\mathrm{F}_0) \approx \mathrm{L}_\mu(\mathrm{F}_0) = 1 - \alpha(\mathrm{F}_0 - \mu) \tag{10}$$

will allow us to search for the optimal parameters α and μ of the function f. The actual slope of the linear approximation to f will vary depending on the coefficients of the linear fits for pitch versus formants of different vowels, as seen in Table I.

In order to simulate the effect of performing VTLN, warp factors were multiplied by the values for the formants of the TIMIT corpus, as measured in Section III. These scaled formants represented the expected result of a corresponding VTLN frequency warping of the input speech signal or features. In order to locate the optimal values for the parameters α and μ, this simulated frequency warping was tested for values of α in the range $[-0.005, 0.005]$ and μ in the range $[50, 250]$.

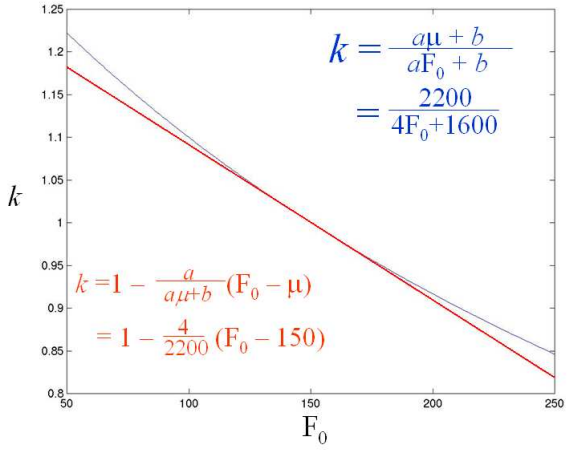

Fig. 3. f and its linear approximation L_μ for $F_{2,/iy/}$

TABLE I

LINEAR FIT COEFFICIENTS AND α FOR FORMANTS OF FOUR VOWELS

Formant	a	b	α
$F_{1,/iy/}$	4.03	1587	1.84×10^{-4}
$F_{2,/iy/}$	0.68	309	1.66×10^{-4}
$F_{1,/ae/}$	0.33	132	1.83×10^{-4}
$F_{2,/ae/}$	0.14	48	2.00×10^{-4}
$F_{1,/aa/}$	0.51	65	3.60×10^{-4}
$F_{2,/aa/}$	0.12	49	1.79×10^{-4}
$F_{1,/uw/}$	0.59	52	4.21×10^{-4}
$F_{2,/uw/}$	0.16	16	3.98×10^{-4}

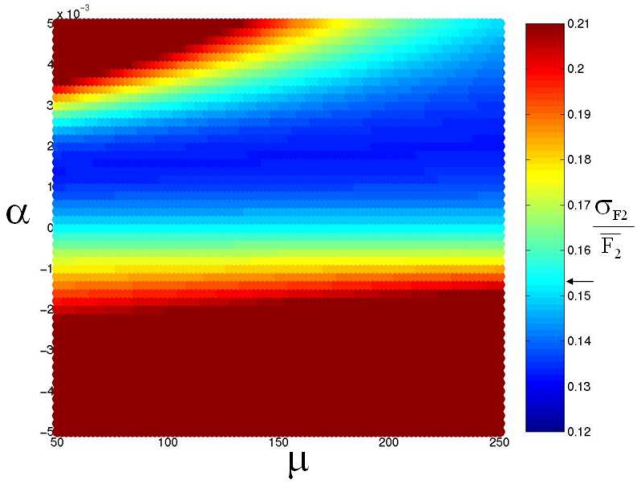

Fig. 4. Normalized standard deviation of warped $F_{2,/iy/}$

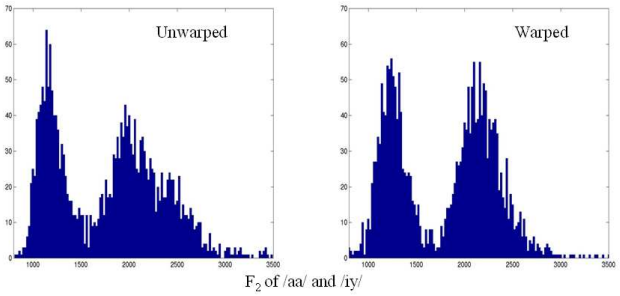

Fig. 5. Distribution of F_2 for the vowels /aa/ and /iy/, before and after a warp determined with $\alpha = 0.002$ and $\mu = 150$

Figure 4 depicts the normalized standard deviation[4] of the values of $F_{2,/iy/}$ when warped by α and μ. It is apparent that an α value of about 0.002 seems to result in the lowest standard deviation of the warped data. An α of 0 corresponds to no warping.

Another measure by which to optimize these parameters is to consider the separability of the data. Consider a binary classification of a signal, such as discriminating the F_2 of the vowels /aa/ and /iy/, as illustrated by the histograms in Figure 5. Borrowing the *d-prime* metric from signal detection theory, the separability of these data may be described by a discriminability index in terms of the means m and standard deviations σ of each class:

$$d' = \frac{|m_1 - m_2|}{\sqrt{\sigma_1 \sigma_2}}$$

Maximizing the discriminability of the formants of two vowels, as reflected by the d-prime scores in Figure 6, it can be seen once again that an α of about 0.002 is optimal.

Observing Figures 4 and 6, there does not appear to be any statistically optimal value for μ[5] . Intuitively, the value for μ should probably be somewhere near the average of the speakers' F_0 values, so as to balance the distribution of warp factors. A suitable value for μ would be in the range from 130 to 170 Hz.

In conclusion, a suitable formula for the warp factors derived by the pitch-based VTLN approach is

$$k = 1 - 0.002(F_0 - 150) \tag{11}$$

[4] Standard deviation divided by mean

[5] ...although the optimal α depends slightly on the selected value for μ, since α is the slope of the linear approximation to (7) at μ.

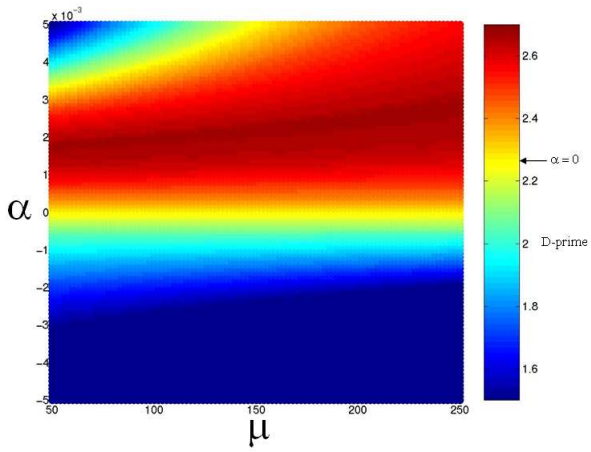

Fig. 6. D-prime scores, indicating discriminability of data

V. Warping functions

The determination of the pitch-based warp factor was the most significant portion of this paper's alternative approach to VTLN. Following that step, there are a number of options for performing the appropriate frequency scaling.

A. Shapes of warping functions

The most straighforward way to implement VTLN is with a linear warping function. The warped frequencies f' are the product of the warp factor k and the original frequencies f:

$$f' = kf$$

Another approach, which accounts for the need to scale higher formants slightly more than lower formants, is a nonlinear warping function, such as described in [5]:

$$f' = k^{\frac{3f}{8000}} f$$

Alternatively, a commonly used warping function is piecewise linear [2]; it is identical to the linear warping function up to a certain frequency f_o, and then the slope is adjusted so as to intersect with the Nyquist frequency f_{N}:

$$f' = \begin{cases} kf & : \quad f \leq f_o \\ kf_o - \frac{f_{\mathrm{N}} - kf_o}{f_{\mathrm{N}} - f_o} & : \quad f_o < f < f_{\mathrm{N}} \end{cases}$$

The piecewise linear warping function ensures more complete usage of the full warped bandwidth. One potential hazard of the linear warping function is that it multiplies by a uniform factor over the entire spectrum, causing considerable changes to the higher frequencies (above the second formant). This region of the spectrum is significant in characterizing fricatives, such as /s/ and /sh/, whose articulation is not at all

dependent on vocal tract length. Applying a large frequency shift to the higher frequencies can increase the confusability of these phones. This effect is especially pertinent for signals sampled at higher sampling rates (such as the TIMIT data at 16 kHz).

B. Two ways of shifting frequency

After determining a warp factor and choosing a type of warping function, the actual frequency shift can be applied to the speech input signal in two manners:

- Resampling or otherwise transforming the original audio signal.
- Adjusting the spacing of filters during the feature calculation stage.

In almost all VTLN implementations, the second of these is utilized. For this project, the principal results were generated using a linear warping function during feature calculation. However, I will also present a different way to achieve the frequency shift, applying time-scaling and resampling prodedures in signal processing prior to feature calculation. This is beneficial in circumstances where the tool for calculating features is a blackbox that does not allow reconfiguration of filter placements.

C. A warping function during pre-processing

Despite an inability to alter the feature calculation stages, I was able to devise a set of processes that effect the equivalent of a linear or piecewise linear warping function by transforming the input audio signal before it is passed on to the feature calculations.

One way to achieve a linear warp in the frequency axis of a signal is to scale it in the time domain; for discrete-time signals, this involves resampling. Given a warp factor k, the warped spectrum $X(\frac{j\omega}{k})$ can be derived via the time-scaling property of the Fourier Transform:

$$x(kt) \overset{FT}{\longleftrightarrow} \frac{1}{|k|} X\left(\frac{j\omega}{k}\right)$$

An unfortunate side-effect of resampling is that the input's time axis is also scaled. Since a great deal of important linguistic information is sensitive to temporal features (see [6]), such as phone duration and voice onset timing, it would be better to retain the original signal's timescale[6].

It is possible to achieve the desired frequency warp without modifying the time axis by utilizing some tools more commonly employed for speech synthesis. The SOLAFS time-scale modification algorithm [7] is an example of a function that can transform a signal's time axis independent of its frequency axis:

$$x(t), X(j\omega) \longrightarrow \boxed{\text{SOLAFS}} \longrightarrow x\left(\frac{t}{k}\right), X(j\omega)$$

[6] Additionally, modifying the time axis could require adjusting transcriptions or phonetic labels, as would be the case with TIMIT.

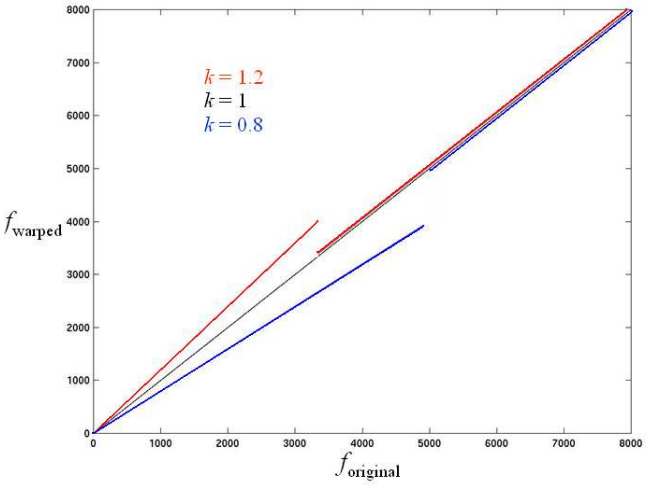

Fig. 7. A warping function realized by time-scaling and resampling a lowpass-filtered portion of a signal and adding it to the highpass-filtered portion of the original signal.

The resulting signal can then be resampled to derive the warped frequency axis along with the original time axis:

$$x\left(\frac{t}{k}\right), X\left(j\omega\right) \longrightarrow \boxed{\text{Resample}} \longrightarrow x'\left(t\right), X\left(\frac{j\omega}{k}\right)$$

Using the above sequence of time-scaling followed by resampling can produce a linear warp of the frequency axis. From this basis it becomes trivial to approximate a piecewise linear warp. First, the signal is split by two filters into a portion containing high frequencies and a portion containing only low frequencies. The low portion undergoes the frequency warp and is then added back to the higher frequencies of the original signal. The cutoff frequency for the lowpass filter should be determined by the warp factor k. For example:

$$f_{cutoff} = \frac{3500}{k}$$

The cutoff for the highpass filter could be similarly scaled, although a fixed value, such as 4000 Hz in this example, might be equally suitable. This piecewise (discontinuous) warping function is shown in Figure 7.

VI. Experimental Design and Results

The pitch-based approach to vocal tract length normalization was implemented and tested to determine the approach's feasibility and relative performance against established VTLN schemes.

A. Comparing warp factors

Evaluation of the pitch-based approach consisted of a comparison of recognizer performance using gender-independent warp factors derived in three ways:

1. **F_0-based** warp factors were calculated for each speaker by using the warp factor formula in (7) and using the speaker's average pitch. The pitch was detected using a fairly reliable pitch-tracking algorithm and the average pitch for a speaker was taken to be the mean of the medians over eight utterances.
2. **Maximum Likelihood** warp factors were calculated while training Gaussian Mixture Models with mel frequency cepstral coefficients (MFCC) features.
3. **F_3-based** warp factors were calculated by tracking the third formant in all voiced frames of eight utterances per speakers. As with the pitch-based warp factors, the speaker's average F_3 was the mean of the medians per utterance. The warp factor was calculated as in (3): the ratio of the speaker's average F_3 to the mean value for the entire set, about 2900 Hz.

These warp factors were linearly warped during the feature calculation stage by the *feacalc* in-house software at the International Computer Science Institute. Using the standard 25ms frames stepped by 10ms, three types of features were generated:

- Mel-PLP: MFCC-like features. Triangular mel-spaced filters; 12 features derived from 22 by DCT truncation.
- PLP-12: trapezoidal bark-spaced filters; 12th order.
- PLP-5: trapezoidal bark-spaced filters; 5th order.

All features included energy plus first-order and second-order derivatives. For each feature, per-utterance normalization shifted the means to zero and variances to one.

An additional set of MFCC features was calculated using tools from the Stanford Research Institute (SRI); these features were identical to the features used for determination of the ML warp factors.

Neural network recognizers were trained on 3696 utterances from 462 speakers in the TIMIT train set. Each neural network was configured with an input layer comprising features with 9 windows of context; the hidden layer had about 164,800 weights[7]; the output consisted of 61 phones. About 10% of the train set was used for cross-validation. The TIMIT test set included 1344 utterances from 168 speakers.

The results are presented as frame accuracy rates, the percentage of phones in a test set that are correctly identified during a forward-pass of a neural net. The results for the experiments with 16 kHz TIMIT data are in Table II; the TIMIT corpus was downsampled to 8kHz and resulting recognition performance is evidenced in Table III. Tests using the MFCC features calculated by SRI's tools [8] are presented in Table IV, which also includes cross-validation rates and evidence of improvement using twice as many hidden units.

In general, the pitch-based warp factors compared quite favorably in relation to the warp factors derived by the formant-based approach. Perhaps this is because pitch is a better indicator of vocal tract length than F_3; more likely, the better performance of the F_0-based warp factors can be attributed to the greater reliability of the pitch-detection tools, as opposed to the formant-tracking software. This is consistent with

[7] The MFCC-like, SRI MFCC, and PLP12 features all comprised 39 features and thus had 400 hidden units. PLP5 had 18 features and 739 hidden units.

[8] The SRI tool also differs from ICSI's *feacalc* in its piecewise linear frequency warping function.

TABLE II

Recognition of test set (frame acurracy rates) with different methods for deriving VTLN warp factors (16kHz)

Feature	Baseline	F_0-based	ML	F_3-based
Mel-PLP	57.29	**59.16**	59.01	57.09
PLP-12	61.74	**61.81**	61.82	60.85
PLP-5	60.19	**60.31**	60.35	60.20

TABLE III

Recognition of test set (frame acurracy rates) with different methods for deriving VTLN warp factors (8 kHz)

Feature	Baseline	F_0-based	ML	F_3-based
Mel-PLP	52.58	**54.03**	55.41	52.17
PLP-12	60.17	**60.21**	60.26	60.14
PLP-5	59.81	**59.95**	59.98	59.83

results independently obtained in [8]. Compared to the ML-based warp factors, however, the pitch-based performance is fairly close.

It is interesting that VTLN seemed to work well with MFCC or MFCC-like features, while there was hardly any difference between warped and unwarped PLP features. This perhaps owes something to differences between triangular and trapezoidal filters used in the respective feature calculations.

It is also interesting or suspicious that the results using SRI features showed such a dramatic difference between training cross-validation rates and test accuracy rates, where the improvement of VTLN is reflected in the CV rates, but not in a forward-pass of the test set. Nonetheless, these results still show that F_0-based VTLN performs very similarly to the ML approach.

TABLE IV

Recognition of test set, as well as CV rates from training, for MFCC features calculated with SRI's tools. An additional set, with 800 hidden units, was also evaluated.

	Baseline	F_0-based	ML	F_3-based
400HU, test	60.73	**60.19**	60.09	60.00
800HU, test	61.41	**60.87**	60.75	60.69
400HU, CV	61.79	**62.41**	62.73	61.17
800HU, CV	62.69	**63.31**	63.44	62.00

TABLE V
BEST CROSS-VALIDATION FRAME ACURRACY RATES (% CORRECT) WITH DIFFERENT WARPING FUNCTIONS (16 KHZ)

Feature	Baseline	Linear	Piecewise Linear
PLP-12	63.09	55.01	56.16
CRB19+d	61.32	52.58	53.47
CRB19+d+dd	61.11	52.35	53.83

B. Comparing warping functions

In addition to looking at results for various approaches to warp factor determination, the effect of a linear versus (discontinuous) piecewise linear warping function was also examined. These warpings were accomplished before features were calculated, using the time-scaling and resampling procedure detailed in the previous section. Three sets of features were calculated: PLP12, as above; CRB19+d, 19 critical band filters plus first-order derivatives; CRB19+d+dd, 19 critical band filters plus first and second-order derivatives.

The results for this experiment are provided in Table V and show that the uniformly linear warping function performs consistently worse than the warping function that splits the audio signal into halves and warps only the lower frequencies. The cross-validation frame accuracies for VTLN are lower than the baselines because there were some signal distortions caused by the SOLAFS algorithm.

These results should be tempered by the fact that observing cross-validation rates is an imperfect way to measure recognizer performance. A better test would have been to connect these phone recognition systems to a language model, allowing a full decoding that returns word error rates.

VII. CONCLUSION

Pitch-based vocal tract length normalization is a new approach to determine warp factors that are experimentally shown to succeed in improving speech recognition performance. At the very least, it seems to give results comparable to using the established ML approach. This could contribute to ASR systems incorporating VTLN as an easy yet effective alternative to existing speaker normalization approaches. Especially when pitch is already being observed – used in prosody-based utterance segmentation, rich transcription, or emotion classification – then the calculation of pitch-based warp factors requires minimal computational expense.

In addition to these good results, there are indications that pitch-based VTLN using a different warping function could yield even greater improvements.

VIII. Acknowledgements

Thanks to Dan Ellis of Columbia University, Andreas Stolcke of ICSI, and David Gelbart of ICSI for their advice and technical expertise.

References

[1] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Atlanta, GA, May 1996, vol. 1, pp. 353–356.

[2] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Atlanta, GA, May 1996, vol. 1, pp. 339–341.

[3] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," in *IEEE Transactions on Speech and Audio Processing*, September 2002, vol. 10, pp. 415–426.

[4] T. Claes, I. Dologlou, L. ten Bosch, and D. Van Compernolle, "A noverl feature transformation for vocal tract length normalization in automatic speech recognition," in *IEEE Transactions on Speech and Audio Processing*, November 1998, vol. 6, pp. 549–557.

[5] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Atlanta, GA, May 1996, vol. 1, pp. 346–348.

[6] H. Hermansky and S. Sharma, "TRAPS – classifiers of temporal patterns," in *Proc. Int. Conf. on Speech and Language Processing*, Sidney Australia, November 1998.

[7] D. Hejna and B. Musicus, "The SOLAFS time-scale modification algorithm," Tech. Rep., Bolt Beranek & Newman, "July" 1991.

[8] C. Lopes and F. Perdig ao, "VTLN through warp factors based on pitch," in *Revista Brasileira de Telecomunicações*, vol. In press; submitted 2003.

[9] Gordon Peterson and Harold Barney, "Control methods in a study of the vowels," *Journal of the Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, March 1952.

[10] D. Whalen and A. Levitt, "The universality of intrinsic F0 of vowels," *Journal of Phonetics*, , no. 23, pp. 349–366–1752, 1995.

Arlo Faria Arlo Faria is an undergraduate at the University of California, Berkeley, where he is progressing towards a degree in Electrical Engineering / Computer Science, as well as minoring in Linguistics. His research interests are in the field of speech and language processing.