Efficient Pitch-based Estimation of VTLN Warp Factors

Arlo Faria and David Gelbart

International Computer Science Institute
1947 Center Street, Suite 600, Berkeley, CA 94704-1198
{arlo,gelbart}@icsi.berkeley.edu

Abstract

To reduce inter-speaker variability, vocal tract length normalization (VTLN) is commonly used to transform acoustic features for automatic speech recognition (ASR). The warp factors used in this process are usually derived by maximum likelihood (ML) estimation, involving an exhaustive search over possible values. We describe an alternative approach: exploit the correlation between a speaker's average pitch and vocal tract length, and model the probability distribution of warp factors conditioned on pitch observations. This can be used directly for warp factor estimation, or as a smoothing prior in combination with ML estimates. Pitch-based warp factor estimation for VTLN is effective and requires relatively little memory and computation. Such an approach is well-suited for environments with constrained resources, or where pitch is already being computed for other purposes.

1. Introduction

Automatic speech recognition systems must be able to cope with considerable variation among speakers; major sources of this inter-speaker acoustic variation are physiological factors such as gender and vocal tract length. VTLN is a technique for scaling the frequency axis of acoustic feature vectors so that observations are more similar across all speakers. This is especially useful in gender-independent systems, since on average the vocal tract is 2-3 cm shorter for females than males, causing females' formant frequencies to be about 15% higher.

The most common method for finding warp factors for VTLN invokes the maximum likelihood (ML) criterion to choose a warp factor that gives a speaker's warped observation vectors the highest probability [1, 2]. The likelihoods can be computed using the recognizer's phone models; alternatively, warp factors can be chosen to maximize likelihoods from reference acoustic Gaussian Mixture Models (GMMs).

Other approaches predict warp factors by observing more direct parameters of speech acoustics, such as formants (resonant frequencies of the vocal tract). The first and second formants can be modeled by vowel-specific distributions [3], or the less phone-dependent third formant can be averaged globally [4]. While these might be good indicators of vocal tract length, accurate formant estimation is difficult – especially in noisy signals. In [5], a warp factor is computed using the ratio of a speaker's pitch to a reference value. We believe this approach is not optimal, since pitch is not directly proportional to vocal tract length. According to [4], formant frequencies are directly proportional to VTL, so if pitch were directly proportional to VTL it would also be directly proportional to F_2, which Figure 1 suggests it is not (note that the axes in Figure 1 do not start at the origin).

This work presents an approach inspired by the correlation between laryngeal size and vocal tract length, as explored in [6]. During training, a joint distribution of pitch and warp factors is estimated by accumulating likelihoods of warped acoustic observations at measured pitch values. This distribution can be utilized to select a most probable VTLN warp factor given a speaker's average pitch, or as a pitch-based prior for combination with the likelihood scores used in ML warp factor estimation.

The process of selecting warp factors can be reduced to pitch extraction, which will generally reduce computation and memory resources needed for VTLN. Using pitch-based warp factors for VTLN provided substantial improvement over a system with no VTLN, and the accuracy approached that of the more computationally intensive ML-estimated warp factors.

2. ML warp factor estimation

For a speaker i, let X_i^α be acoustic observation vectors with a frequency axis scaled by warp factor α. Given the corresponding transcriptions, the acoustic data likelihood could be computed using an HMM acoustic model for Viterbi alignment; but since these transcriptions are unavailable during testing, they must be hypothesized from a prior decoding pass [1]. Alternatively, a mixture of multivariate Gaussians (GMM) can be used to model generic speech frames, enabling warp factor selection to be moved entirely into the front-end processing [2].

Given observed data X_i and a reference GMM acoustic model λ, the probability of a warp factor α can be described in terms of acoustic likelihoods:

$$\mathrm{P}(\alpha|X_i,\lambda) = \frac{P(X_i^\alpha|\lambda)}{\sum_{\alpha'} P(X_i^{\alpha'}|\lambda)} \qquad (1)$$

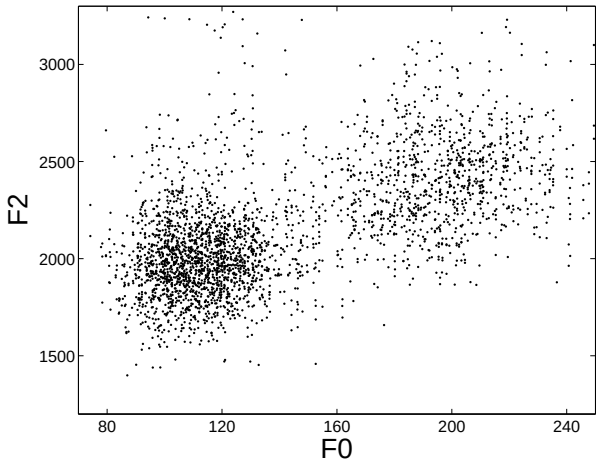

Figure 1: Frequencies of the second formant F_2 plotted against a speaker's average pitch F_0, for all segments of the vowel /iy/ in the TIMIT corpus. Data from [6].

The optimal warp factor is selected by searching over a range of α values:

$$\hat{\alpha}_i = \arg\max_{\alpha} \mathrm{P}(\alpha|X_i, \lambda) \tag{2}$$

This maximizes the likelihood of the warped utterances X_i^{α}, which is desirable because this same criterion is used for MLE model training and in recognition.

3. Pitch-based warp factor estimation

There is a correlation between a speaker's average pitch and the vocal tract resonances. In general, both are dependent on the physical size and gender of the speaker [7]. For example, a large male is generally larger in all dimensions, and tends to have not only a longer vocal tract but also a wider glottis and thicker vocal folds. This can greatly affect speech characteristics: Figure 1 illustrates this correlation, plotting the second formant for a given vowel segment in relation to the speaker's average pitch.

Pitch-based warp factor estimation simply requires a conditional probability $\mathrm{P}(\alpha|\mathrm{F}_0)$. We associate a single value of F_0 to each speaker: f_i is considered the mean pitch over speaker i's voiced frames. Given the observed pitch $\mathrm{F}_0 = f_i$, a speaker's optimal warp factor is found:

$$\hat{\alpha}_i = \arg\max_{\alpha} \mathrm{P}(\alpha|f_i) \tag{3}$$

3.1. Modeling $\mathbf{P}(\alpha|\mathbf{F}_0)$ from training data

To calculate the conditional probability of α given an observed mean pitch f_i, we utilize the joint distribution:

$$\mathrm{P}(\alpha|f_i) = \frac{\mathrm{P}(f_i, \alpha)}{\sum_{\alpha'} \mathrm{P}(f_i, \alpha')} \tag{4}$$

The joint probability can be approximated during training by summing indicator functions to get counts:

$$\mathrm{P}(f, \alpha) \approx \frac{\sum_i I_i(f, \alpha)}{\sum_{i, f', \alpha'} I_i(f', \alpha')} \tag{5}$$

A simple procedure counts one (f, α) observation per speaker. Unlike pitch, the speaker's warp factor is not directly observable from data; we might use the warp factor $\hat{\alpha}_i$ selected by the ML methods in the previous section. Then a joint observation would be counted as

$$I_i(f, \alpha) = \begin{cases} 1, & \text{if } f = f_i \text{ and } \alpha = \hat{\alpha}_i \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

In practice, the training data is too sparse to construct a smooth distribution using just one joint observation per speaker. So we choose a more robust solution, basing counts on the observation likelihoods of a speaker's warped utterances. Using the GMM likelihoods, from Eq. (1):

$$I_i(f, \alpha) = \begin{cases} \mathrm{P}(\alpha|X_i, \lambda), & \text{if } f = f_i \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

Figure 2 depicts the conditional probability of warp factors given pitch (Eq. 4) which was trained for our experiments[1]. For smoothness, a zero-phase ten-point moving average filter was applied along the F_0 dimension.

3.2. Combination of pitch-based and ML estimates

We also try combining Eqs. (2) and (3), where a pitch-based prior allows warp factors to be selected with a maximum *a posteriori* criterion:

$$\hat{\alpha}_i = \arg\max_{\alpha} \ \mathrm{P}(\alpha|X_i, \lambda) \cdot \mathrm{P}(\alpha|f_i) \tag{8}$$

Note that the terms are unweighted; it would also be possible to set interpolation weights using held-out data.

[1] We work with systems that define warp factors inverse to the standard convention, scaling filterbank frequencies by $1/\alpha$.

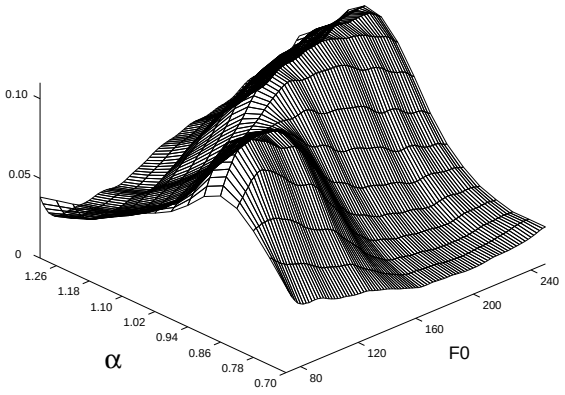

Figure 2: The conditional probability $P(\alpha|F_0)$.

4. Experiments and results

Experiments were devised to compare the performance of systems using no VTLN, ML warp factor estimation, and pitch-based warp factor estimation. The OGI Numbers95 corpus, with a vocabulary of 80 words, was suitable for these tests due to the wide range of speakers; about 3000 training speakers (3 hours) and 1000 test speakers (1 hour) were selected from the corpus.

These systems were based on SRI's DECIPHER recognizer [8], where 39 mel-cepstral features were processed with mean and variance cepstral normalization (CN). Feature warping for VTLN was implemented with a piecewise linear scaling of the filterbank frequencies.

Transcripts of the training set were used to estimate a bigram language model, and the decoder was run in a one-pass configuration generating 1-best hypotheses. All systems described in this paper were gender-independent.

Table 1 displays the results of these experiments. To observe the effects due to the amount of speech data used in normalization, we tried VTLN and CN in both a per-utterance and per-speaker scheme (respectively, an average of 1.6 s and 3.3 s of data). The baseline systems in the first row used only CN and no VTLN (i.e. $\alpha = 1$).

For pitch-based VTLN (labeled 'Pitch-based' in Table 1), we used the ESPS get_f0 program [9]. We quantized F_0 as 251 values from 50 to 300 Hz, and warp factors as 16 values from 0.70 to 1.30. Pitch-based warp factors were selected as in Eq. (3).

A contrastive system used DECIPHER's ML warp factor estimation, calculating acoustic likelihoods with a frame-level Gaussian mixture model (as in [2]). The reference GMM was iteratively estimated from the training set, and warp factors were considered in the same range as with pitch-based estimation.

Finally, we tried a system (labeled 'Combined' in Table 1) that combined pitch-based and ML estimates, as described in Section 3.2.

Warp factor estimation	Normalized per-utterance	Normalized per-speaker
None	6.5	5.6
Pitch-based	5.2	4.7
ML	5.0	4.5
Combined	4.7	4.5

Table 1: Experimental results (word error rate, in %)

5. Discussion

5.1. Improvement in recognition accuracy

The results in Table 1 show that pitch-based warp factor estimation can be an effective method of improving ASR performance, as there is a substantial reduction in word error rate compared to a system with no VTLN. Thus pitch was useful for estimating warp factors, even when pitch was averaged over fairly short utterances (the Numbers95 task does not elicit much intra-speaker prosodic pitch variation, which may have been helpful in this regard). The performance of the purely pitch-based approach was almost as good as that of the usual ML method for warp factor estimation.

The combined approach appeared to give improved results over ML alone, but only when normalization was performed on an utterance – rather than speaker – basis; this is intuitive if we consider pitch information in the combined approach to be a prior which becomes less useful when there is more data available to the ML estimation. (Consider Figure 3, which plots acoustic likelihoods from the reference GMM used for ML warp factor estimation. Among utterances from a particular speaker, these likelihoods differed enough such that the optimal warp factors would vary considerably. Using that speaker's three utterances together, the per-speaker likelihoods became less noisy.) The results for the combined approach suggest it may be useful for systems in which VTLN is to be performed with limited amounts of speech data.

5.2. Resource usage and implementation costs

There is considerable demand for ASR on platforms with limited available memory and computing power, which motivates our interest in reducing the computation and memory required for VTLN.

For pitch-based estimation, computing the warp factor involves little more than pitch extraction and a table lookup; in our experiments, this proceeded nearly five times faster than computing likelihoods of warped utterances for the ML approach. Disregarding algorithmic changes which trade thoroughness for speed (such as a golden-section search, or grid search over smaller ranges), we compared against one of the most efficient ML estimation procedures. So it is plausible that pitch-based estimation is the faster approach, generally. Fur-

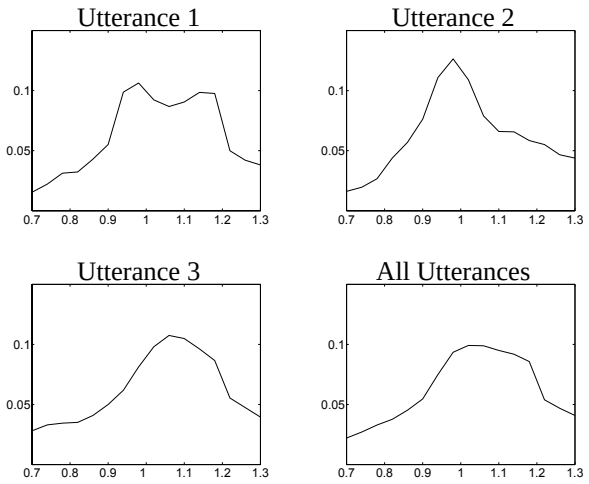

Figure 3: $P(\alpha|X_i, \lambda)$ over a range of warp factors. The first three plots are for individual utterances from a single speaker. The bottom-right corresponds to that same speaker's three utterances considered in aggregate.

thermore, if the recognition system already performs pitch extraction for other purposes (e.g., for noise-robust feature extraction [10] or to exploit prosody [11]), then pitch-based warp factors enable VTLN at a trivial cost.

The memory requirements for pitch-based warp factor estimation are also small. Whereas an ML system may require storage of a reference acoustic model for calculating warp factors (DECIPHER implements a 128-Gaussian mixture model), a pitch-based system only requires storage of the most probable warp factor for each pitch. This relation could also be approximated by a linear regression: in previous work [6] we estimated warp factors as a function of pitch, with a best-fit line characterized by just two parameters: slope and intercept.

6. Conclusion

This paper presents an effective procedure for VTLN warp factor estimation, exploiting the correlation between pitch and vocal tract length. The reduced resource requirements of this novel approach make it an appealing alternative for VTLN on constrained architectures. Our work also suggests that a pitch-based prior can be used to improve ML warp factors estimated from scarce data.

We have created a webpage accompanying this paper, which provides Matlab code and additional discussion, and may be used for new information in the future: www.icsi.berkeley.edu/Speech/papers/eurospeech05-vtln

7. Acknowledgements

Thanks to Jeremy Ang, Barry Chen, Horacio Franco, Michael Shire, and Andreas Stolcke for their input. Arlo Faria was supported by UC Berkeley's SUPERB undergraduate research program and the European Union 6th FWP IST Integrated Project AMI (Augmented Multi-party Interaction, FP6-506811). David Gelbart was supported by the German Ministry for Education and Research's SmartWeb project and by a Canadian NSERC fellowship.

8. References

[1] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *ICASSP*, Atlanta, GA, May 1996, vol. 1.

[2] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *ICASSP*, Atlanta, GA, May 1996, vol. 1.

[3] M. Lincoln, S.J. Cox, and S. Ringland, "A fast method of speaker normalisation using formant estimation," in *Eurospeech*, Rhodes, 1997.

[4] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *ICASSP*, Atlanta, GA, May 1996, vol. 1.

[5] C. Lopes and F. Perdigão, "VTLN through warp factors based on pitch," *Revista Brasileira de Telecomunicações*, vol. 18, no. 1, June 2003.

[6] A. Faria, "Pitch-based vocal tract length normalization," Tech. Rep. TR-03-001, International Computer Science Institute, 2003.

[7] D. R. Smith, R. D. Patterson, R. Turner, H. Kawahara, and T. Irino, "The processing and perception of size information in speech sounds," *Journal of the Acoustical Society of America*, vol. 117, no. 1, January 2005.

[8] A. Stolcke et al., "The SRI March 2000 Hub-5 conversational speech transcription system," in *Proc. NIST Speech Transcription Workshop*, University of Maryland, May 2000.

[9] "Open source software from the speech group," *www.speech.kth.se/software/*.

[10] M. L. Seltzer, J. Droppo, and A. Acero, "A harmonic-model-based front end for robust speech recognition," in *EUROSPEECH*, Geneva, Switzerland, 2003.

[11] E. Shriberg and A. Stolcke, "Direct modeling of prosody: An overview of applications in ASR," in *Proc. International Conference on Speech Prosody*, Nara, Japan, 2004.