On Line Vocal Tract Length Estimation for Speaker Normalization in Speech Recognition

William R. Rodríguez, Oscar Saz, Antonio Miguel and Eduardo Lleida

Communications Technology Group (GTC)
Aragón Institute for Engineering Research (I3A)
University of Zaragoza, Zaragoza, Spain
{wricardo,oskarsaz,amiguel,lleida}@unizar.es

Abstract

This paper presents the results on an Automatic Speech Recognition (ASR) framework that takes advantage of robust vocal tract length estimation methods for improving the performance of speech recognition in the presence of speakers with different conditions in age and gender. Well known techniques for Vocal Tract Length Normalization (VTLN) usually require previous stages for the estimation of the best warping factor for a given speaker, either by Maximum Likelihood (ML) estimates or by the calculation of acoustic features from the speakers like formant frecuencies through several utterances. This paper will show how to use robust framewise estimations of the vocal tract length to obtain a speaker dependent warping factor for achieving major improvements over all conditions of the TIDigits database. In the end, an updating function will be used to calculate an on-line estimate of the vocal tract length and the warping factor to use real time VTLN in speech recognition with similar results to the off-line strategies.

Index Terms: vocal tract length estimation, speech recognition, speaker normalization

1. Introduction

The mismatch between the set of speakers used to train a given acoustic model for Automatic Speech Recognition (ASR) and the set of speakers which are recognized in that ASR system can seriously degrade the performance of the recognition results. A well known source of mismatch are the anatomical features that different speakers may have regarding the structure or their vocal tract. The vocal tract, and more precisely its length, varies largely from one speaker to another, especially if the range of speakers gathers males, females, adults or children. All this can make an ASR system trained on adults perform poorly in the presence of children speech and vice versa.

Different possibilities have arisen for the reduction of this mismatch between training data and recognition data. Some of them require a re-training of the acoustic models, as in speaker adaptation techniques like Maximum A Posteriori (MAP) [1] or Maximum Likelihood Linear Regression (MLLR) [2]; while some others act on the speech signal and keep the models unchanged. Vocal Tract Length Normalization (VTLN) is a well known technique for reducing this mismatch without modifying the initial acoustic modeling [3]. It considers that the main difference between two speakers is the change in the frequency axis due to the difference in vocal tract length between the speakers.

However, VTLN techniques usually require of large computational delays, as they need to process speech data from the speaker in advance to estimate which is the best transformation from the speaker's frequency axis to the target speakers' frequency axis. This makes that most of VTLN-based techniques can not provide their improvement in the ASR performance in a real-time situation. The proposal in this work wants to advance in the field of providing this improvement in a real-time on-line framework. It makes use of robust speech processing algorithms to give a frame-by-frame estimation of the vocal tract length of the speaker, in such a way that it allows for providing a transformation factor for a given speaker without requiring more information that the current frame and previous frames.

The paper is organized as follows: Section 2 will review the basis of VTLN techniques and the different existing approaches. Section 3 will present the signal processing techniques which lead to a robust framewise vocal tract length calculation for all speakers and Section 4 will present the three VTLN techniques evaluated in this paper, including the on-line real-time approach. Next, in Section 5 the results and improvements achieved with the evaluated methods over the different conditions of the TIDigits database will be shown. Finally, Section 6 will provide the conclusions to this work.

2. Vocal Tract Length Normalization

The aim of VTLN is to provide a warping function that transforms the frequency axis from a given speaker (f) to the frequency axis of a target speaker or a target group (f'). Many different possibilities have been researched in the literature to provide the function which reflects this transformation, from piecewise linear approaches to exponential functions. All of them depend on a warping factor, α like in Equation 1, which contracts or expands the spectrum of the speech signal in the desired way [3].

$$S_{warped}(f) = S_{unwarped}(f'(\alpha, f)) \qquad (1)$$

A warping factor that contracts the frequency axis is used to transform speakers with shorter vocal tracts (mainly children and women) towards speakers with the longest vocal tracts (i.e. men), and a warping factor that expands this axis is used for warping speakers with longer vocal tracts (men) towards the shortest ones (children or women). A more efficient proposal has been proposed for ASR consisting in transforming and warping the Mel-scale filter banks during the Mel-Frequency

This work was supported by national project TIN2008-06856-C05-04 from the Spanish government and Santander Bank scholarships

Cepstral Coefficients (MFCC) calculation, instead of warping all the input frames from the speaker.

2.1. Estimation of the Warping Factor

The estimation of the warping factor α for a given speaker or utterance is the most delicate part in the use of VTLN. An inadequate factor may reduce the potential improvement provided by VTLN, or even produce a loss in performance.

Two trends in the proposals for estimating the warping factor can be observed in the literature: On one hand, Maximum Likelihood (ML) based proposals select that warping factor which achieves the highest score by forced aligning several versions of the input utterances, warped with different factors, to the acoustic model for recognition [3]; on the other hand, feature-based proposals use acoustic features from the speaker like formants, or a combination of them, to estimate the warping factor, as it is known that the formant frequencies correlate to the vocal tract length of the speaker [4].

3. Vocal Tract Length Estimation

Although many methods rely on the estimation of speaker features which can be correlated to the vocal tract length to calculate the optimum warping factor, there have been little efforts to estimate the actual vocal tract length. Difficulties in the estimation of this anatomical measure, especially in the presence of voices with a high fundamental frequency, have limited the development of methods based on direct vocal tract length estimation. This Section will describe a robust method to estimate this value for all possible speakers with the aim of using it as estimator of the optimal warping factor.

Modeling the vocal tract as a uniform lossless acoustic tube, its resonant frequencies given by Equation (2) are uniformly spaced, where $v = 35300$ cm/s is the speed of sound at 35 °C, and l is the length of the uniform tube in cm.

$$F_k = \frac{v}{4l}(2k-1), k = 1, 2, 3, \ldots \quad (2)$$

The estimation of the length was proposed in [5], and it can be reduced to fitting the set of resonance frequencies of a uniform tube, which are determined solely by its length l. Therefore, the problem can be approximated to minimizing Equation 3, where $D(\tilde{F}_k, (2k-1)F1)$ is a function that express the difference between the measured formants $\tilde{F}_k(k = 1, ..., M)$ and the resonance of the uniform tube.

$$\varepsilon = \sum_{k=1}^{M} D(\tilde{F}_k, (2k-1)F_1) = \sum_{k=1}^{M} D(\tilde{F}_k, (2k-1)\frac{v}{4l}) \quad (3)$$

From [5], the error measure given in equation (3) can be turned in Equation 4 using the distance function between the measured formants$(\tilde{F}_k)$ and the odd resonances of a uniform tube, $(2k-1)F_1$.

$$\varepsilon = \sum_{k=1}^{M} \frac{(\frac{\tilde{F}_k}{2k-1} - F_1)^2}{F_1} \quad (4)$$

The formant frequencies $\tilde{F}_k$ are extracted using traditional Linear Prediction Coefficients (LPC) method with order $p = 8$, over a 25 ms long speech frame. The filter coefficients for the all-pole vocal tract model are obtained through Durbin's recursion using the autocorrelation method, after Hamming-windowing the pre-emphasized speech frame.

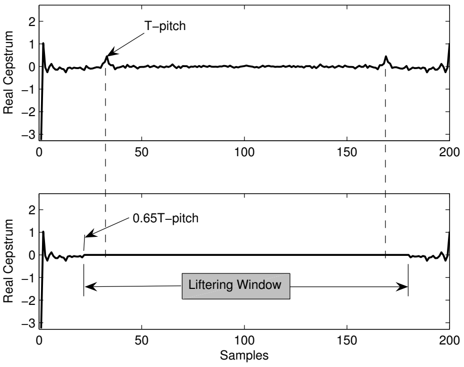

Figure 1: *Effect of liftering in the real cepstrum domain*

Finally, the vocal tract length can be obtained with the expression in Equation 5 which makes use of the estimated resonance frequency of the uniform tube (F_1), calculated from Equation 4 as in Equation 6.

$$VTL = \frac{v}{4F_1} \quad (5)$$

$$F_1 = \left(\frac{1}{M} \sum_k \left(\frac{\tilde{F}_k}{2k-1} \right)^2 \right)^{1/2} \quad (6)$$

3.1. Robust Formant Estimation in High Pitch Voices

The formant measurement is technically difficult. The situation is less severe in male adult cases in which the fundamental frequencies (F0) are low [6]. In women and children $F0$ increases, so $F0$ and its harmonics could get closer to the range of the formant values affecting the estimation [7]. The conventional autocorrelation method with the LPC parameters works well in signals with a long pitch period (low-pitched), but as the pitch period in high-pitched speech is small, the periodic replicas cause aliasing in the autocorrelation sequence. In that case it is required to separate these effects in order to obtain formants not contaminated by $F0$ by means of homomorphic analysis.

The main idea within the homomorphic analysis is the deconvolution of a segment of speech $x[n]$ into a component representing the vocal tract impulse response $e[n]$, and a component representing the excitation source $h[n]$ as in Equation 7.

$$x[n] = e[n] * h[n] \quad (7)$$

The way in which such separation is achieved is through linear filtering of the cepstrum, defined as the inverse Fourier transform of the log spectrum of the signal. As the cepstrum in the complex domain is not suitable to be used because of its high sensitivity to phase[8], the real-domain cepstrum $c[n]$ defined by Equation (8) is used, where $X(k)$ is the N-point Fourier transform of the speech signal $x[n]$.

$$c[n] = \frac{1}{N} \sum_{k=0}^{N-1} ln\left|X(k)\right| e^{j\frac{2\pi}{N}kn}, 0 \leq n \leq N-1 \quad (8)$$

The values of $c[n]$ around the origin correspond primarily to the vocal tract impulse information, while the farthest values are affected mostly by the excitation. Knowing previously the value of the pitch period T_{pitch} from the LPC analysis using the autocorrelation method it is possible to filter the cepstrum signal (liftering) and use the liftered signal to find the formant frequencies, once the signal is back in the time-domain.

A liftering window with the length of $0.5T_{pitch}$ has been proposed in [9] or $0.6 - 0.7T_{pitch}$ in [10]. In this work, the liftering window $w[n]$ is $0.65T_{pitch}$ and the effect of appliying $w[n]$ in the real cepstrum domain can be observed in Figure 1.

4. VTLN Techniques Applied in ASR

Two VTLN techniques were initially compared in this work, first one was a state of the art ML-based approach, while second one was based on the proposal for robust vocal tract length calculation in Section 3. An exponential function was used for the warping of the Mel-scale bank filters in the MFCC calculation. No model adaptation was performed in the experiments and, hence, no Jacobian compensation was done as previous results reported how this feature degraded performance in the presence of unadapted models [11, 12].

The ML-based technique was based on the diagram seen in Figure 2 [3], where an initial ASR stage obtained an estimate of the transcription of the uttered sentence. A set of n Viterbi alignment decoders using a set of warping factors $\{\alpha_1 ... \alpha_n\}$ decided the most likely of those warping factors according to the score achieved by each decoder. Finally, that warping factor was used in a second ASR stage which made use of VTLN to improve the estimation of the output utterance and provide a final result. This work used 11 warping factors in the Viterbi decoding phase, ranging from 0.9 to 1.1 in 0.02 intervals.

The feature-based technique to be evaluated in this work used the framewise estimation of the vocal tract length as seen in Section 3. This estimation provided a value of the vocal tract length in the sonorant frames and a void output in the rest of frames (silence and unvoiced sounds). From all the valid estimations of one speaker, the mean of the vocal tract length for the speaker was calculated (VTL_{spk}) and the warping factor was obtained as in Equation 9, where $\overline{VTL}_{model}$ was the mean of the vocal tract lengths calculated for all the speakers used in the training of the acoustic model, which could be easily done in the prior training phase. The factor λ was used to moderate the amount of warping applied, and was set to $\lambda = 0.5$ after a prior development set of experiments on smaller databases.

$$\alpha = 1 + \lambda \frac{\overline{VTL}_{model} - VTL_{spk}}{\overline{VTL}_{model}} \tag{9}$$

4.1. On-line Vocal Tract Length Estimation

The main drawback of the VTLN techniques shown previously was that they required previous stages to estimate the warping factor. The ML-based strategy required three stages for each utterance (initial recognition, decision of the warping factor and final recognition), while the feature-based strategy required that several utterances from a speaker were available to estimate robustly the mean of the vocal tract length from that speaker. These approaches were not feasible for real world applications which should provide a real-time decoding of the input speech utterance.

The proposal in this work was, as the calculation of the vocal tract length was robust in a framewise approach, to re-estimate the vocal tract length for each frame as in Equation 10, where β is the memory factor of the system. The value estimated for the vocal tract length in a given frame i only depended on its value in the previous frame $i - 1$ and the actual value of vocal tract length estimated for frame i: $VTL(i)$. This approach avoided the influence of local variations of the vocal tract length, while tending to the mean value of the speaker when sufficient frames were analyzed. A memory factor of $\beta = 0.99$ was used for the experimentation.

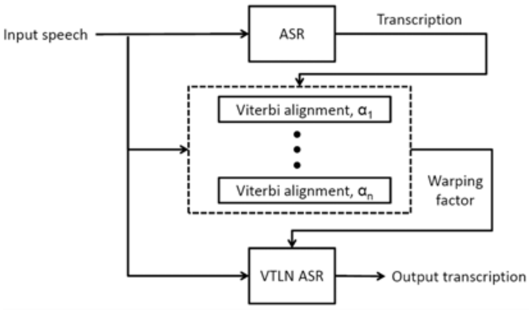

Figure 2: *ML-based VTLN diagram*

$$VTL_{spk}(i) = \beta * VTL_{spk}(i-1) + (1-\beta) * VTL(i) \tag{10}$$

This way, when a speaker accessed for the first time to the ASR system, the vocal tract length was initialized with the vocal tract length of the target model as in Equation 11; as every new sonorant frame was available and a value of the vocal tract length was provided for that frame, the vocal tract length for the speaker was updated according to Equation 10 (again with $\lambda = 0.5$), and the warping factor for that frame calculated as in Equation 12.

$$VTL_{spk}(0) = \overline{VTL}_{model} \tag{11}$$

$$\alpha(i) = 1 + \lambda \frac{\overline{VTL}_{model} - VTL_{spk}(i)}{\overline{VTL}_{model}} \tag{12}$$

5. Experimental Framework and Results

The evaluation of the techniques proposed here was made over the TIDigits database [13]. This corpus contains 25 boys, 26 girls, 55 men and 57 women for the training of models and 25 boys, 25 girls, 56 men and 57 women for the recognition evaluation. Seven conditions were designed with seven different acoustic models trained for each condition: Boys, girls, men, women, adults (men and women), child (boys and girls) and all speakers. Recognition was performed over all the 163 speakers available for evaluation.

A set of 11 word Hidden Markov Models (HMM) representing digits in English were trained for each condition. An ETSI-like front end was used to extract the MFCC parameters from each signal, using the 12 first static parameters ($c1$-$c12$) plus log-energy and their first and second derivatives for the final 39 dimension feature vectors. The ASR system used for the experiments was a state of the art Viterbi decoder.

The baseline results in first row of Table 1 in terms of Word Error Rate (WER) showed the big influence of acoustic mismatch between models and speakers for recognition. The recognition results provided are the ones obtained for the recognition of the test data from boys, girls, men and women. Worst results were achieved with models from men speech, as men presented the longest vocal tracts, which separates greatly from the rest of the speakers. On the other edge, girls have the shortest vocal

Table 1: *Results in WER for the TIDigits database: Models trained on boy, girl, man, woman, adult, child and all speech*

	Boy	Girl	Man	Woman	Adult	Child	All
Baseline	7.37%	19.21%	25.17%	5.32%	2.01%	8.20%	0.65%
Off-line ML-based VTLN	2.47%	5.26%	8.58%	1.28%	1.05%	2.40%	0.57%
Off-line vocal tract estimated VTLN	2.84%	5.37%	11.25%	1.94%	1.19%	2.81%	0.66%
Off-line vocal tract estimated VTLN (liftering)	2.35%	3.92%	10.15%	1.57%	1.07%	2.18%	0.65%
On-line vocal tract estimated VTLN	2.61%	4.78%	10.48%	1.82%	1.18%	2.49%	0.65%

Table 2: *Mean vocal tract length in cms with standard deviation intervals estimated for the speaker groups in the TIDigits database*

Train speakers				
	Boy	Girl	Man	Woman
VTL	16.0±0.64	15.5±0.65	18.8±0.67	16.6±0.64
Test speakers				
	Boy	Girl	Man	Woman
VTL	15.9±0.74	15.4±0.58	18.8±0.71	16.6±0.63

tract and the models trained from girls speech also performed poorly. The model trained with all the speech matched perfectly the recognition speakers and achieved a 0.65% of WER.

The off-line techniques to be evaluated in this work achieved the results in the second, third and fourth rows of Table 1 for the ML-based VTLN and two versions of the feature-based technique respectively. The performance of both techniques was similar, with some differences across all the conditions which indicated that the vocal tract length estimation in Section 3 was as good as the state of the art techniques for speaker normalization in ASR. It was noticed the improvement that the use of liftering, seen in Section 3.1, produced in reduction of the WER; showing the need for using robust formant estimation techniques when dealing with variable speech.

More precisely, the VTLN based on direct estimation of the vocal tract length with liftering achieved better results with those models trained on boys, girls and children altogether; while the ML-based technique performed better with models trained on men, women and adults. Performance on the model trained with all speakers was similar for both techniques. Table 2 shows the mean values with their standard deviation of the vocal tract lengths for all the speakers in the TIDigits database with the estimation method of Section 3. These values confirm the big mismatch between the different group, especially men versus the rest of the groups; and confirmed the need of applying the speaker normalization techniques studied in this work. The robustness of the estimation of these values was seen in their good statistical properties across speakers.

Finally, the on-line technique proposed in Section 4.1 achieved the results shown in fifth row of Table 1. Although a certain decrease of performance was noticed throughout most of the conditions, the results were comparable to those of the off-line version of the algorithm and confirmed the possibilities of performing real-time VTLN in ASR with results similar to the state of the art off-line techniques.

6. Conclusions

The main result and conclusion of this work is the development of a method for applying successful on-line speaker normalization in ASR. This method relies in a robust estimation of the vocal tract length from the speaker at the frame level, which allows to apply a warping factor which can be updated and improved as more data from the user is available. This overcomes the drawback from traditional VTLN techniques where speech utterances are processed in several stages to obtain the desired improvement. Furthermore, the proposed method relies only in voice features from the speaker and is not based in models and likelihood measures as previous approaches.

The use of techniques for enhancing the formant estimation methods and thus, improving the vocal tract length calculation, makes suitable this speaker normalization method for all types of speakers (children or adults), independently of the value in fundamental frequency of the speaker. This is a relevant improvement over other techniques for the estimation of acoustic features like formants which may face difficulties in the presence of speech with a high fundamental frequency.

7. References

[1] J.-L. Gauvain and C.-H. Lee, "Maximum A Posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[2] C.-J. Legetter and P.-C. Woodland, "Maximum Likelihood Linear Regression for speaker adaptation of the parameters of continous density Hidden Markov Models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[3] L. Lee and R. Rose, "A frequency warping approach to speaker normalization," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 6, no. 1, pp. 49–60, 1998.

[4] E.-B. Gouvea and R.-M. Stern, "Speaker normalization through formant-based warping of the frequency scale," in *Proceedings of Eurospeech*, Rhodes, Greece, 1997, pp. 1139–1142.

[5] F.-N. Burhan, A.-C. Mark, and P.-B. Thomas, "Unsupervised estimation of the human vocal tract length over sentence level utterances," in *Proceedings of ICASSP'00*, 2000, pp. 1319–1322.

[6] H. Traunmuller and A. Eriksson, "A method of measuring formant frequencies at high fundamental frequencies," in *Proceedings of Eurospeech)*, 1997, pp. 477–480.

[7] W. Rodríguez and E. Lleida, "Formant estimation in children's speech and its application for a spanish speech therapy tool," in *Proceedings of SLaTE*, Wroxhall Abbey, UK, 2009.

[8] R. Schafer and L. Rabiner, *Digital Processing of Speech Signals*. Prentice-Hall, 1978.

[9] W. Verhelst and O. Steenhaut, "A new model for the short-time complex cepstrum of voiced speech," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 2, pp. 43–51, 1986.

[10] M. Shahidur and T. Shimamura, "Formant frequency estimation of high-pitched speech by homomorphic prediction," *Acoustic Sci. and Tech.*, vol. 26, no. 6, pp. 502–510, June 2005.

[11] A.-N. Harish, D.-R. Sanand, and S. Umesh, "Characterizing speaker variability using spectral envelopes of vowel sounds," in *Proceedings of Interspeech'07*, Brighton, UK, 2009, pp. 1107–1110.

[12] S. Panchapagesan and A. Alwan, "Frequency warping for vtln and speaker adaptation by linear transformation of standard mfcc," *Computer, Speech and Language*, vol. 1, no. 23, pp. 42–64, 2009.

[13] R.-G. Leonard, "A database for speaker independent digit recognition," in *Proceedings of ICASSP'84*, San Diego, CA (USA), 1984, pp. 328–331.