Real-Time Vocal Tract Length Normalization in a Phonological Awareness Teaching System

Dénes Paczolay[1], András Kocsor[2], and László Tóth[3]

Research Group on Artificial Intelligence
of the Hungarian Academy of Sciences and University of Szeged*
H-6720 Szeged, Aradi vértanúk tere 1., Hungary
{[1]pdenes, [2]kocsor, [3]tothl}@inf.u-szeged.hu
http://www.inf.u-szeged.hu/speech

Abstract. Speaker normalization in a speech recognition can significantly improve speech recognition accuracy. One such method, vocal tract length normalization (VTLN), is especially useful when the system has to work reliably for males, females and children. It is just this situation with our phonological awareness teaching system, the "SpeechMaster", which aims at real-time phoneme recognition and feedback. As most VTLN algorithms work off-line, this poses the additional problem of real-time operation. This paper examines how a well-known off-line algorithm can be approximated on-line by machine learning regression techniques. We conclude that, by employing a real-time estimation of VTLN parameters, the recognition error can be reduced by some 14-24 %.

1 Motivation

Real-time phoneme recognition – combined with a real-time visualization of the results – forms the basis of many phonological awareness drilling systems such as in the "SpeechMaster" software developed by our team. Because the robustness & reliability of the recognizer is a vital issue of these systems, techniques like real-time speaker normalization examined in this paper are of great importance. By means of off-line vocal tract length normalization (VTLN) algorithms [2–4, 6–9] one can build recognizers that work robustly with male, female and children's voices. However, we were confronted with the problem that the normalization parameters had to be estimated on-line. To solve the problem we used nonlinear regression methods of machine learning. With their application our on-line approximations closely approach the recognition results of the off-line methods.

The structure of the paper is as follows. In the next section we present the "SpeechMaster" program designed for teaching reading and speech therapy. Section 3 describes how to perform speaker normalization via (off-line) VTLN, while Section 4 discusses the approximation of the off-line parameters by on-line regression. Lastly, Section 5 presents our experimental findings along with a brief discussion and conclusion.

* This work was supported under the contract IKTA No. 2001/055 from the Hungarian Ministry of Education.

(a) (b)

Fig. 1. The user interface of the "SpeechMaster" software. (a-b) The teaching reading part and the speech therapy part, respectively.

2 A Phonological Awareness Teaching System

The "SpeechMaster" software developed by our team seeks to apply speech recognition technology to speech therapy and the teaching of reading. The role of speech recognition is to provide a visual phonetic feedback. In the first case it is to replace the missing auditive feedback of the hearing impaired, while in the case of the latter it is to reinforce the correct association between the phoneme-grapheme pairs. With the aid of a computer children can practice without the need for the continuous presence of the teacher. This is very important because the therapy of the hearing impaired requires a long and tedious fixation phase. Furthermore, experience shows that most children prefer computer exercises to conventional drills.

Both applications require a real-time response from the system in the form of an easily comprehensible visual feedback. With the simplest display setting it is given by means of flickering letters, their identity and brightness being adjusted to the speech recognizer's output. Figure 1 shows the user interface of "SpeechMaster", in the teaching reading and the speech therapy applications, respectively. As one can see, in the first case the flickering letter is positioned over a traditional picture for associating the word and word sound, while in the latter case it is combined with a web camera image that helps the impaired student learn the proper articulator positions.

Since the system should work reliably both for children and teachers of different ages, the recognizer has to be trained with the voices of users of both genders and of practically any age. The task is also special because it has to recognize isolated phones, so it cannot rely on language models. Consequently, there is a heavy burden on the acoustic classifier, and we need to apply any helpful trick we can think of. One such technique is speaker normalization, or more specifically, vocal tract length normalization (VTLN), which proves very useful when the targeted users vary greatly in age and gender.

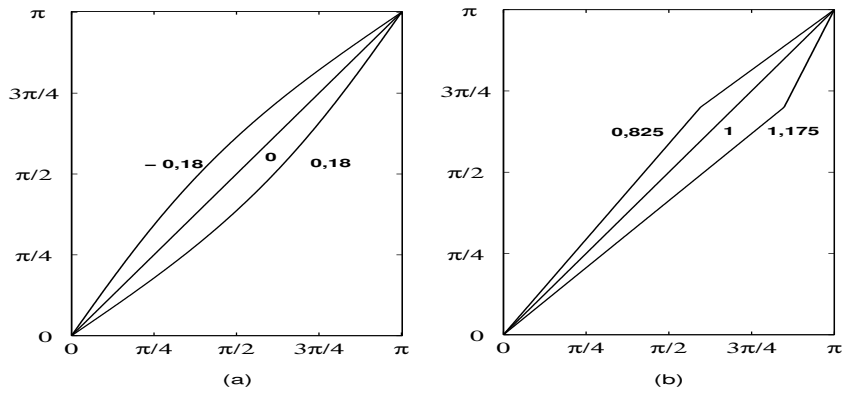

Fig. 2. Two examples of VTLN frequency warping functions: (a) The bilinear warping function, and (b) a piecewise linear warping function.

3 Speaker Normalization using VTLN

One of the major physiological sources of inter-speaker variation is the vocal tract length of the speakers. In [2] the average vocal tract length for men was reported to be 17 cm, for women it was 15 cm, and for children it was 14 cm. The formant frequency positions are inversely proportional to vocal tract length and this causes a shift of the formant center frequencies. Consequently, VTLN is usually performed by warping the frequency scale. Modelling the vocal tract as a uniform tube of length L, the format frequencies are proportional to 1/L. Thus the simplest approaches use a linear warp. In reality, however, the vocal tract is more complex than a uniform tube. That is why many more sophisticated warping functions have been proposed in the literature [3, 7]. Two of these, the bilinear function and a piecewise warping function, are shown in Fig. 2, and are given by the following formulas:

$$f' = \arctan \frac{(1-\alpha^2)\sin f}{(1+\alpha^2)\sin f - 2\alpha} \qquad \alpha \in [-0.18, 0.18], \tag{1}$$

$$f' = \begin{cases} \alpha f & \text{if } 0 < f < 0.7/\alpha \\ \gamma f + (1-\gamma) & \text{otherwise.} \end{cases} \quad \alpha \in [0.825, 1.175], \quad \gamma = \frac{0.3\alpha}{\alpha - 0.7}. \tag{2}$$

Given a warping function, normalization can be implemented either by re-sampling and interpolating the spectrum or modifying the width and center frequencies of the mel (Bark) filter bank.

Experience shows that the critical issue is not how the warping function is parameterized but, rather, how the optimal parameter for a given user is obtained. Some of these techniques are described in [6, 8, 9]. A common feature of these methods is that they usually require at least a whole utterance for optimal performance. In our case, however, instantaneous adjustment is needed. Thus in the following we study the viability of on-line parameter tuning.

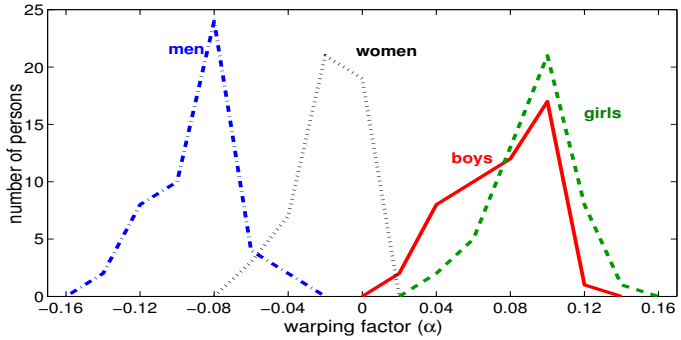

Fig. 3. Distribution of the optimal warping factor for the bilinear warping function, as found by LD-VTLN.

4 Real-Time Estimation of VTLN Parameters

Having chosen a warping function we need to estimate the optimal set of warping parameters for each speaker of a given speech database based on a proper optimality criterion. Although this criterion varies from method to method (cf. [3, 7, 8]), their comparison is not in the scope of this paper. Instead we restrict our investigations to the linear discriminant based criterion suggested in [8]. This method was reported to be just as efficient but more stable than the ubiquitous maximum likelihood approach [3].

First, we will briefly summarize the main steps of the LD-VTLN method [8]. The linear discriminant (LD) criterion is defined using the covariance matrices of a given sample set over a speech database. Each sample is placed in a phonetic class and the samples that belong to a given speaker are extracted using the same warping parameter. The task is to optimize these parameters for each speaker according to the LD criterion: $LD = |T|/|W|$, where T is the total covariance matrix of all samples and W is the average inter-class covariance matrix. The value of the LD criterion is large if the different classes are spaced out and each of them has a small scatter around the class centers. Since optimizing the warping parameters of all the speakers at the same time is impractical, the following iterative process can be applied:

0. Choose an initial warping factor for each speaker and warp the samples
1. For each speaker
   a. calculate the LD criterion for each α value in a small neighborhood of the current warping parameter
   b. store the best warping parameter
2. Update the sample set using the optimal warping factors obtained
3. If the average warping factor variation is above the set threshold go to 1.

This optimization method, however, works off-line. So a natural question arises. Is it possible to efficiently estimate the optimal parameters obtained by off-line algorithms using machine learning regression methods that work on-line?

To answer this question, out of the many possible regression techniques we chose to experiment with neural nets. Their task was to estimate the optimal LD-VTLN warping parameter for each speaker based on the actual spectral frame without warping. Our specific question was how close this on-line method could come to the off-line LD-VTLN scores. The answer to this is in the next section.

5 Experiments and Evaluation

Firstly we describe the corpus, the feature extraction technique, the classifiers and regression algorithms used in the tests. Then we present the details of the LD-VTLN and the on-line regression experiments.

Corpus. For training and testing purposes we recorded samples from 200 speakers, namely 50 women, 50 men, 50 girls and 50 boys. The children were aged between 6 and 9. The speech signals were recorded and stored at a sampling rate of 22050 Hz in 16-bit quality. Each speaker uttered all the Hungarian vowels, one after the other, separated by a short pause. Since we decided not to discriminate their long and short versions, we only worked with 9 vowels altogether.

Feature Sets. The signals were processed in 10 ms frames, the log-energies of 24 critical-bands being extracted using FFT and triangular weighting [5]. The energy of each frame was normalized separately, which means that only the spectral shape was used for classification.

Classifiers. In all the classification experiments the Artificial Neural Nets (ANN) [1] employed were the well-known three-layer feed-forward MLP networks trained with the backpropagation learning rule. The number of hidden neurons was equal to 16.

Regression. For the learning of the parameter α of the warping function a special MLP network was constructed with one output neuron and two hidden layers with 24 and 24 neurons, respectively. Training was performed with respect to mean square error.

LD-VTLN. The initial value of the warping parameter α was set to 0 for eq. (1) and to 1 for eq. (2). The interval of the possible α values is shown in Fig. 2. For the optimization the value of α in this interval was quantized – it could take one of 15 discrete values. The iteration was stopped when the average change in the warping parameter fell below 10^{-2}. Fig. 3 shows the distribution of the warping parameters obtained from the LD-VTLN algorithm, when running on the full database using the warping function defined in eq. (1).

Tests. The experiment were conducted as follows. The corpus of 200 speakers was downsampled to 3 sets of 50, 100 and 200 speakers, keeping the uniform ratio of boys, girls, men and women in each case. All three databases were divided into train and test sets with a ratio of 80/20%. In the train set we extract the features for all the possible α values, which served as input for the LD-VTLN algorithm. The optimal warping factors found by LD-VTLN along with the original spectral features formed the input for the neural net regression technique. Lastly, neural net classification tests were run on the following settings: a given database without warping (NO VTLN), warping by LD-VTLN (LD-VTLN), and warping via the real-time, regression-based VTLN (Realtime(RT)-VTLN). The results are summarized in Table 1.

	50 speakers			100 speakers			200 speakers		
	NO-VTLN	LD-VTLN	RT-VTLN	NO VTLN	LD-VTLN	RT-VTLN	NO-VTLN	LD-VTLN	RT-VTLN
bilinear	18.52	13.07	14.09	15.36	12.02	13.02	14.33	11.02	11.52
piecewise linear	18.52	13.86	14.35	15.36	12.17	13.19	14.33	10.87	11.04

Table 1. Recognition errors for the vowel classification task. The rows correspond to the two warping functions, while the columns correspond to the database sizes and the normalization methods applied.

Discussion. Looking at Table 1 it can be seen that the (off-line) LD-VTLN method can reduce recognition error by as much as 21-29%. Using the on-line regression estimations we obtained error reductions of about 14-24%, which is very close to the LD-VTLN score. Increasing the size of the database, the discrepancy between the scores of the methods applied diminished. In addition, we saw no significant difference between the two warping functions. In conclusion, we can state that it is worth experimenting with the regression-based estimation of off-line techniques like LD-VTLN because the results are very close and the regression method allows for on-line speaker normalization.

Lastly, we should mention that in phonological awareness development systems the sharp separation of the phoneme classes is almost as important as the recognition score. Speaker normalization techniques – like LD-VTLN and RT-VTLN – are also effective in dealing with this problem. A closer investigation of how these work in practice will be a subject of future work.

References

1. Bishop, C. M., *Neural Networks for Pattern Recognition,* Oxford University Press, 1995.
2. Claes, T., Dologlou, I., Bosch, L. and Compernolle, D., A Novel Feature Transformation for Vocal Tract Length Normalization in Automatic Speech Recognition, *IEEE Trans. on Speech and Audio Processing*, vol. 6, pp. 549-557, 1998.
3. Eide, E., Gish, H., A Parametric Approach to Vocal Tract Length Normalization, *Proc. ICASSP '97,* pp. 1039-1042, Munich, Germany, 1997.
4. Pitz, P., Molau, S., Schlter, R., Ney, H., Vocal Tract Normalization Equals Linear Transformation in Cepstral Space, *Proc. EUROSPEECH-2001,* Vol. 4, pp. 2653 - 2656, Denmark, 2001.
5. Rabiner, L. R., Juang, B.-H., *Fundamentals of Speech Recognition,* Englewood Cliffs, NJ, Prentice Hall, 1993.
6. Uebel, L. F., Woodland, P. C., An Investigation into Vocal Tract Lenght Normalisation, *Proc. EUROSPEECH-99,* Vol. 6, pp 2527 - 2530, Hungary, 1999.
7. Wegmann, S., McAllaster, D.,Orloff, J., Peskin, B. Speaker Normalization on Coversational Telephone Speech, *Proc. ICASSP-96,* Vol. 1, pp. 339-341, Atlanta, 1996.
8. Westphal, M., Schultz, T.,Waibel, A., Linear Discriminant - A New Criterion For Speaker Normalization, *Proc. ICSLP '98,* paper no. 755, Sydney, Australia, 1998.
9. Zhan, P., Westphal, M., Speaker Normalization based on Frequency Warping, *Proc. ICASSP-97,* Vol. 1, pp. 1039-1042, Munich, 1997.