論文 / 著書情報
Article / Book Information

Title	Rapid Vocal Tract Length Normalization using Maximum Likelihood Estimation
Authors	T. Emori, K. Shinoda
Citation	Proc. EuroSpeech2001, Vol. , No. , pp. 1649-1652
Pub. date	2001,

Rapid Vocal Tract Length Normalization using Maximum Likelihood Estimation

Tadashi Emori and Koichi Shinoda

Computer & Communication Media Research, NEC Corporation,
4-1-1 Miyazaki, Miyamae-ku, Kawasaki, 216-8555, Japan
{t-emori,shinoda}@ccm.cl.nec.co.jp

Abstract

Recently, vocal tract length normalization (VTLN) techniques have been developed for speaker normalization in speech recognition. This paper proposes a new VTLN method, in which the vocal tract length is normalized in the cepstrum space by means of linear mapping whose parameter is derived using maximum-likelihood estimation. The computational costs of this method are much lower than that of such conventional methods as ML-VTLN, in which the parameter for mapping is selected from among several parameters. Further, the new method offers greater precision in determining parameters for individual speakers. Experimental use of the method resulted in an error reduction rate of 7.1%. A combination of the proposed method with cepstrum mean normalization (CMN) method was also examined and found to reduce the error rate even more, by 14.6%.

1. Introduction

Extensive studies have been conducted on speaker-independent (SI) speech recognition systems using hidden Markov models (HMMs). Parameters of the HMMs in these systems are estimated using speech data that has already been collected from large numbers of speakers, thus making it unnecessary for new users to enroll in the system. It is well known, however, that the recognition accuracy of such SI systems is usually inferior to that of speaker-dependent (SD) speech recognition systems, in which each user enrolls in the system, and it is especially low for certain individual speakers. Speaker adaptation or speaker normalization has been used in many systems to try to overcome this problem.

In speaker adaptation, the model parameters are re-estimated for each individual user on the basis of only a few utterances from that user. In many speaker adaptation techniques, a mapping from the parameters of the SI model to those of the user's model is created. One such technique is maximum likelihood linear regression (MLLR) [1], in which an affine transformation in the parameter space is estimated.

In speaker normalization, features that are dependent on individual speakers are subtracted from observed features. Extensive studies have been conducted on cepstrum mean normalization (CMN) [2] and vocal tract length normalization (VTLN) [3].

In CMN, the long-term average of the cepstrum is subtracted from the cepstrum of each data frame. This helps eliminate changes created not only by differences among individual speakers, but also by environmental noise and channel changes, for which changes are much slower than the changing phonetic features of speech itself.

Since vocal tract length differs from speaker to speaker, so do the formant frequencies in the power spectrum for each speaker. In VTLN, vocal tract lengths are estimated using each speaker's spectrum, but since it is difficult to precisely estimate a vocal tract length from a spectrum, some studies have used a maximum-likelihood VTLN (ML-VTLN) selection method [4, 5, 6, 7]. With this method, a number of parameters, each of which represents a different vocal tract length, are prepared beforehand, and the parameter that maximizes the likelihood of the data is selected. This method is, however, computationally very costly. Roughly speaking, the computational cost is n times higher than the cost of processes not using this method, where n is the number of parameters prepared, because likelihoods must to be calculated for each parameter. When the number of parameters is decreased to reduce the computational costs, the precision of the parameter estimation deteriorates because smaller numbers of parameters are less capable of representing the wide variety of the vocal tract lengths found among large numbers of speakers.

A method called speaker adaptive training (SAT) has recently come into frequent use [8, 9, 10, 11]. Here, so long as speaker adaptation will always be carried out for each speaker, a standard-speaker-dependent model (i.e., a speaker-dependent model based on the speech of a *standard* speaker) will be more appropriate for use as the initial model than a speaker-independent model (i.e., a model representing variations in the utterances of a large number of speakers). In SAT, the parameters for a standard-speaker-dependent model are estimated in the following process. First, a mapping from the parameters of the model created for each individual speaker to those of an initial model is estimated. Second, this estimated mapping is used to map the utterance data for each speaker. Third, this mapped data is used to train the standard-speaker-dependent model. This process is iterated until convergence. While an affine transformation is often used for the mapping, since the number of parameters to be estimated is relatively large, it is difficult to precisely estimate its parameters when the number of utterances is small.

Both speaker normalization and SAT remove the variations in input speech data caused by differences in individual speaker characteristics. The effectiveness of these methods depends mainly on the method used to conduct the mapping, for which precise estimation needs to be achieved on the basis of only a small amount of data from each speaker.

In this paper we propose a VTLN method that employs linear mapping requiring only a single parameter, one estimated with a maximum-likelihood method. The computational cost is much lower than that for ML-VTLN because it is unnecessary to calculate likelihoods for multiple parameters. Further, the amount of data necessary for estimating the single parameter is

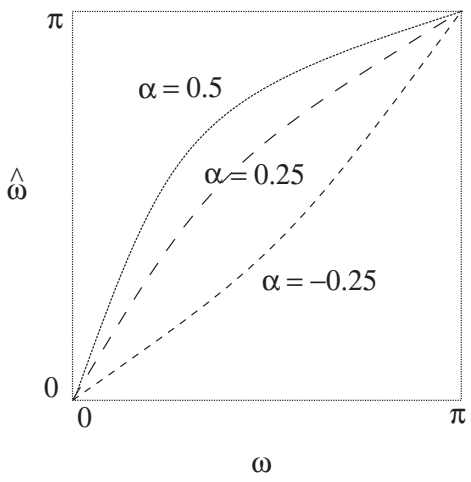

Figure 1: Frequency warping functions

much less than that needed for estimating multiple parameters in SAT that employs affine transformations.

In the next section, we describe the algorithm used in the proposed method, and in Section 3, we discuss the results of our experimental evaluations.

2. Rapid Vocal Tract Length Normalization

2.1. Warping function

While difference between an individual VTL and a standard VTL can be easily represented by a warping function in spectrum space, it is less straightforward task to do so in cepstrum space, which is what is ordinary used in the training and recognition processes in speech recognition. We employ an all-pass transform (1) [12, 13, 14] as a warping function here because its parameter can be estimated in the cepstrum space (as will be explained later in 2.2). The all-path transform is expressed as

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \qquad z = e^{j\omega}, \qquad \hat{z} = e^{j\hat{\omega}}, \tag{1}$$

where α is a warping parameter and $\alpha < 1$; ω and $\hat{\omega}$ are the respective frequencies before and after the transformation. Figure 1 shows examples of warping functions.

2.2. Cepstrum transformation

We will now explain how to calculate the cepstrum coefficients transformed by the warping function. Let c be the cepstrum before the transformation, and $\hat{c}$ be the cepstrum after the transformation. Then, the z-transforms of the cepstrum are:

$$S(z) = \sum_{m=0}^{\infty} c_m z^{-m}, \qquad \hat{S}(\hat{z}) = \sum_{m'=0}^{\infty} \hat{c}_{m'} \hat{z}^{-m'}. \tag{2}$$

If it is assumed that $S(z) = \hat{S}(\hat{z})$ for all z, then

$$\hat{c}_n = \sum_{m=0}^{\infty} c_m \frac{1}{2\pi j} \oint z^{-m} \hat{z}^{n-1} d\hat{z}. \tag{3}$$

From (1) and (3), we have

$$\begin{aligned}
\hat{c}_0 &= \sum_{m=0}^{\infty} \alpha^m c_m \\
\hat{c}_1 &= (1-\alpha^2) \sum_{m=1}^{\infty} m \alpha^{m-1} c_m \\
\hat{c}_2 &= c_2 + \alpha(-c_1 + 3c_3) \\
&\quad + \alpha^2(-4c_2 + 6c_4) \cdots \\
\hat{c}_3 &= c_3 + \alpha(-2c_2 + 4c_4) \\
&\quad + \alpha^2(c_1 - 9c_3 + 10c_5) \cdots \\
&\vdots
\end{aligned} \tag{4}$$

These equations can be rewritten as

$$\hat{\mathbf{c}} = \mathbf{A}_0 \mathbf{c}, \tag{5}$$

where

$$\begin{aligned}
\hat{\mathbf{c}} &= \left(\hat{c}_0\, \hat{c}_1\, \hat{c}_2\, \hat{c}_3 \cdots\right)^t \\
\mathbf{A}_0 &= \begin{pmatrix}
1 & \alpha & \alpha^2 & \alpha^3 \cdots \\
0 & 1-\alpha^2 & 2\alpha - 2\alpha^3 & \cdots \cdots \\
0 & -\alpha + \alpha^3 & 1 - 4\alpha^2 + 3\alpha^4 & \cdots \cdots \\
\vdots & \vdots & \vdots & \vdots\ \vdots \\
\vdots & \vdots & \vdots & \vdots\ \vdots
\end{pmatrix} \\
\mathbf{c} &= \left(c_0\, c_1\, c_2\, c_3 \cdots\right)^t .
\end{aligned} \tag{6}$$

2.3. Estimation of the warping parameter

Next, the warping parameter α is estimated using an maximum-likelihood (ML) estimation.

Let $O = \mathbf{o}_1, \ldots, \mathbf{o}_T$ be a sequence of observed feature vectors used for the estimation, $\mathbf{q} = q_1, \ldots, q_T$ be a state sequence, and $\Theta = (\lambda, \alpha)$ be the parameter set to be estimated, where λ is the HMM parameter set. Then, the auxiliary function to be maximized in the E-M algorithm is set to be

$$Q(\Theta', \Theta) = \sum_{\mathbf{q}} P(\mathbf{O}, \mathbf{q}|\Theta') \log P(\mathbf{O}, \mathbf{q}|\Theta), \tag{7}$$

where Θ' is the current estimate of the parameter set Θ. When it is assumed that the output probabilistic density function (pdf) for each HMM state is a Gaussian pdf with a diagonal covariance and the HMM parameter set λ is unchanged, (7) can be rewritten as

$$Q(\alpha', \alpha) = \sum_{j=1}^{J} \sum_{t=1}^{T} \gamma_t(j) \log b_j(\hat{\mathbf{c}}_t), \tag{8}$$

$$b_j(\hat{\mathbf{c}}_t) = \frac{1}{\sqrt{(2\pi)^M |\mathbf{\Sigma}_j|}} \exp\left[-\frac{1}{2}(\hat{\mathbf{c}}_t - \mu_j)^t \mathbf{\Sigma}_j^{-1} (\hat{\mathbf{c}}_t - \mu_j)\right], \tag{9}$$

where J is the number of HMM states, M is the dimension of feature vectors, μ_j is the mean vector at state j, Σ_j is the diagonal covariance at state j, in which σ_{mj} is the m-th diagonal element, and $\gamma_t(j)$ is the posterior probability of being in state j at time t.

By differentiating the right-hand side of (8) by α and setting its result equal to 0, we have

$$\sum_{j=1}^{J} \sum_{t=1}^{T} \frac{\gamma_j(t) \frac{\partial b_j(\hat{\mathbf{c}}_t)}{\partial \alpha}}{b_j(\hat{\mathbf{c}}_t)} = 0. \tag{10}$$

It is impossible to analytically solve this equation with respect to α, since it is a polynomial function of α. While McDonough et al. [12, 13] solve this equation by using Newton's method, the acquired α is not always the optimal one. Here, it is assumed that the HMM parameter set is trained using the utterances from many speakers, and therefore, the warping parameter α is sufficiently small for unknown speakers, $\alpha \ll 1$. Under this assumption, the transform matrix $\mathbf{A}_0$ in (5) is approximated as

$$\mathbf{A} \equiv \begin{pmatrix} 1 & \alpha & 0 & 0 & \cdots \\ 0 & 1 & 2\alpha & 0 & \cdots \\ 0 & -\alpha & 1 & 3\alpha & \cdots \\ 0 & 0 & -2\alpha & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix}, \tag{11}$$

in which α^n with $n > 1$ is ignored. Finally, we have

$$\alpha = \frac{\sum_{j=1}^{J}\sum_{t=1}^{T}\gamma_t(j)\left[\sum_{m=1}^{M}\frac{1}{\sigma_{mj}^2}\Delta c_{mjt}\bar{c}_{mt}\right]}{\sum_{j=1}^{J}\sum_{t=1}^{T}\gamma_t(j)\left[\sum_{m=1}^{M}\frac{1}{\sigma_{mj}^2}\bar{c}_{mt}^2\right]}, \tag{12}$$

where c_{mt} is the m-th cepstrum coefficient at time t, and

$$\begin{aligned} \Delta c_{mjt} &= c_{mt} - \mu_{jm}, \\ \bar{c}_{mt} &= (m-1)c_{(m-1)t} - (m+1)c_{(m+1)t}. \end{aligned} \tag{13}$$

We refer to this algorithm as rapid VTLN (R-VTLN). This algorithm can easily be applied to HMMs with Gaussian-mixture output pdfs. So as you can see, the computational cost for obtaining the warping parameter is very low.

2.4. Application to HMM recognition

This R-VTLN can easily be combined with the HMM training and recognition process. In training, the warping parameter α is first fixed, and then the HMM parameter set λ is estimated. Next, λ is fixed and α is estimated. This process is iterated until convergence. In recognition, a small number of user utterances are used to estimate α while λ is fixed.

To enhance the effect of R-VTLN, the algorithm is modified for the following two reasons. First, it can generally be assumed that the cepstrum coefficients in the lower order are more significant in the estimation of the warping parameter α. Therefore, only the cepstrum coefficients of the order less than a certain threshold are used for estimation, which means that the threshold should be experimentally optimized. Second, the estimation of α using consonants is much less reliable than that made using vowels. Therefore, only the speech data for the vowel phones are used to make the estimation. It should be noted that once the warping parameter is estimated, the same warping parameter is applied to all of the cepstrum coefficients for all of the phones in the recognition process.

3. Experiments

3.1. Experimental conditions

The proposed method was evaluated using Japanese isolated-word recognition. Each utterance was digitized at a sampling rate of 11.025 kHz, and analyzed in 16-ms frame periods. The frequency range was 300–5000 Hz. The analysis yielded a vector with 21 components (a power derivative, 10 mel-scaled cepstrum coefficients, and 10 corresponding mel-scaled time derivatives). We used the demi-syllable [16] as the recognition unit. The number of states in each unit was three, and the number of mixture components in each state was two. A diagonal covariance was used for each mixture component.

In the training, we used speech data from 56 speakers. Each of these speakers uttered 2,000 phonetically-balanced words. In the test, we used 100 city names as speech data. Each word was uttered twice by 90 speakers (45 male and 45 female); the first time for the estimation, and the second time for recognition. In the recognition experiments, we used a vocabulary of 5,000 words which including the 100 city names. It was assumed that the transcription of each utterance was known in the estimation.

Table 1: Results of recognition experiments (%)

	Male	Female	Ave.
SI	78.7	78.8	78.8
ML-VTLN	79.4	79.1	79.3
R-VTLN(1)	79.4	79.0	79.2
R-VTLN(2)	80.1	80.4	80.3
CMN	80.1	81.4	80.7
CMN+R-VTLN(2)	81.4	82.3	81.9

3.2. Experimental results

The proposed method was first compared with conventional methods. The results of the recognition experiments are shown in Table 1. The results of the speaker-independent recognition experiment are shown in the SI category. For ML-VTLN [7], several warping parameters were prepared beforehand, and the parameter that maximized the likelihood of the data for estimation was chosen. In this experiment, the transformation in (5) was used as the warping function, and thirteen warping parameters, from -0.3 to +0.3 in 0.05 intervals, were prepared beforehand. We conducted the experiment by using two versions of the proposed method: R-VTLN(1) and R-VTLN(2). While all 10 of the cepstrum coefficients for all the phones were used for the estimation of α in R-VTLN(1), the two refinements described in 2.4 were implemented in R-VTLN(2). The order of the cepstrum coefficients used for the estimation was optimized after conducting several preliminary experiments and set to four (from the first to the fourth). As shown in Table 1, R-VTLN(1) achieved a recognition accuracy nearly equal to that of ML-VTLN. It was proven that R-VTLN accurately estimated the warping parameter at a lower computational cost than that of ML-VTLN. In addition, use of R-VTLN(2) reduced the error rate by 5.3% in comparison to R-VTLN(1). Thus the two refinements were proven to be effective. The total error reduction rate of R-VTLN(2) in comparison to SI was 7.1%. This proves the effectiveness of the proposed method.

Table 1 shows the results of two other experiments, CMN and CMN+R-VTLN(2). In CMN, the recognition experiment was carried out sequentially utterance by utterance. For each utterance, the cepstrum mean was updated by averaging the cepstrums for all of the past utterances, and was subtracted from the cepstrum of each frame. In CMN+R-VTLN(2), CMN was followed by R-VTLN(2). Experimental results showed that CMN+R-VTLN(2) was more accurate than CMN. The error reduction rate was 6.2%. These results showed that the gain obtained by the proposed method and that obtained by CMN were independent; the combination of these two method show bet-

Table 2: Word accuracy obtained with a small amount of adaptation data (%)

No.Words	SI	1	2	3	4	5
Male	80.0	80.5	81.4	81.1	81.4	81.1
Female	77.1	79.8	79.6	79.3	79.5	79.7
Ave.	78.5	80.2	80.5	80.2	80.5	80.4

ter performance than either of them when used separately. The overall error reduction rate of the combined method in comparison to SI was 14.6%.

Next, we determined how much data was necessary to estimate the warping parameter. In this experiment, the amount of data varied between one utterance and five utterances, and the five words corresponding to these utterances were excluded from the recognition vocabulary list. The results of the recognition experiment using R-VTLN(2) are shown in Table 2. While the estimation using four utterances was most accurate, the estimation using only one utterance had nearly the same accuracy. It was thus proven that in the proposed method, the parameter could be efficiently estimated even with an extremely small amount of data.

4. Conclusions

We have described a new VTLN method, which we refer to as R-VTLN. In this method, the warping parameter in spectrum space is estimated using the maximum likelihood estimation in cepstrum space. The error reduction rate of the proposed method was 7.1%, and that of the combination of the proposed method and CMN was 14.6%. The method achieved higher recognition accuracy than ML-VTLN and had a much lower computational cost.

This method can be regarded as an SAT method that uses a linear mapping with only one parameter. Only one utterance is required to estimate the parameters, which is much less than the amount needed to estimate the parameters for the SAT that uses affine transformation. Therefore, this method is significantly effective even when the number of user utterances available is extremely small.

In the future studies, this method should be evaluated using speech data from speakers for whom recognition accuracy is extremely low in speaker-independent recognition. It should also be examined under conditions in which environmental noise is significantly high. The method should also be evaluated for cases in which the transcription of the user's utterances is not available.

5. References

[1] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous-density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.

[2] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304-1312, 1974.

[3] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. ICASSP96*, vol. 1, pp. 346-3483, 1996.

[4] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. ICASSP96*, vol. 1, pp. 353-356, 1996.

[5] S. Wegmann, D. Macllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. ICASSP96*, vol. 1, pp. 339-341, 1996.

[6] L. Welling, S. Kanthak, and H. Ney, "Improved methods for vocal tract normalization," in *Proc. ICASSP99*, no. 1436, 1999.

[7] P. Zhan, M. Westohal, "Speaker normalization based on frequency warping," in *Proc. ICASSP97*, pp.1039-1042, 1997.

[8] T. Anastasakos, J. McDonough, R. Schwartz, and John Makhoul, "A compact model for speaker-adaptive training," in *Proc. ICSLP96*, vol. 2, FrP2L1.3, 1996.

[9] H. Jin, S. Matsoukas, R. Schwartz, and F. Kubaka, "Fast robust inverse transform speaker adapted training using diagonal transformations," in *Proc. ICASSP98*, vol. 2, pp. 785-788, 1997.

[10] D. Pye and P .C. Woodland, "Experiments in speaker normalization and adaptation for large vocabulary speech recognition," in *Proc. ICASSP97*, vol. 2, pp. 1047-1050, 1997.

[11] L. Welling, R. Haeb-Umbach, X. Aubert, and N. Harberland, "A study on speaker normalization using vocal tract normalization and speaker adaptive training," in *Proc. ICASSP98*, vol. 2 , pp. 797-800, 1998.

[12] J. McDonough and W. Byrne, "Speaker adaptation with all-path transforms," in *Proc. ICASSP99*, no. 2093, 1999.

[13] J. McDonough and W. Byrne, "Single-pass adapted training with all-pass transforms," in *Proc. EUROSPEECH99*, vol. 6, pp. 2737-2740, 1999.

[14] A. V. Oppenheim and D. H. Johnson, "Discrete representation of signals," in *Proc. IEEE*, vol. 60, pp. 681-691, 1972.

[15] A. V. Oppenheim and R. W. Shafer, *Discrete-Time Signal Processing*, Prentice-Hall, 1989.

[16] K. Yoshida, T. Watanabe, and S. Koga, "Large vocabulary word recognition based on demi-syllable hidden Markov model using small amount of training data," in *Proc. ICASSP-89*, pp. 1-4, 1989.