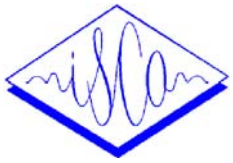


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

MODEL-BASED SPEAKER NORMALIZATION METHODS FOR SPEECH RECOGNITION

Masaki Naito[1], *Li Deng*[2], *Yoshinori Sagisaka*[1]

[1]ATR Interpreting Telecommunications Research Labs., 2–2, Hikaridai, Seika-cho,
Soraku-gun, Kyoto, 619-02 Japan
[2]Department of Electrical and Computer Engineering, University of Waterloo,
Waterloo, Ontario, Canada N2L 3G1

ABSTRACT

We have developed a speaker normalization method using model-based vocal-tract (VT) geometric parameters. The nonlinear frequency warping functions are estimated based on the formant frequencies of each speaker's crude VT geometry approximated by modifying the reference speaker's VT tube and used to normalize the acoustic properties of speakers. This method further provides detailed frequency warping functions specific to individual phonemes by manipulating parameters in the VT model. The results of the phoneme recognition experiments show that our new speaker normalization method is superior in performance to the conventional data-driven speaker adaptation and normalization methods while drastically reducing the amount of adaptation data needed to estimate the speaker normalization parameters.

10.21437/Eurospeech.1999-633

1. INTRODUCTION

Recently, many data-driven speaker adaptation and normalization methods have been proposed, but they have required a large amount of adaptation data to improve performance of speaker-independent speech recognition. In contrast, the acoustic differences between speakers reflect the vocal-tract (VT) geometric differences in a highly nonlinear and indirect fashion. We believe that the various acoustic features of each speaker's speech can be computed and modified by taking into account some simple vocal-tract geometric parameters associated with the speaker, and that such VT parameters can be used to advantage for speaker adaptation and normalization. In this paper, we propose a new speaker normalization method for speaker-independent speech recognition based on using the VT geometric parameters and the related VT model. In this method, we first approximate the VT model of the target speaker based on two factors: 1) phoneme-dependent factor: A standard area functions that depend on phonemes and is estimated by manipulating an articulatory model; and 2) speaker dependent factor: Two dimensional VT geometric parameters, including the length of the oral section (l_1) of the VT and the pharyngeal section (l_2) of the VT, that can be estimated from a small amount of speech data[3]. Based on these factors, the speaker and phoneme-dependent VT shapes are approximated by stretching or shrinking the standard area function according to the l_1 and l_2 parameters of the target speaker. Then the acoustic features of each speaker's speech are computed by taking into account the resonance frequencies of the approximated VT shapes. In the work reported in this paper, we apply the new acoustic features obtained above to estimate the phoneme-dependent frequency warping function for speaker normalization. The frequency warping functions are determined from the relationship between the VT resonance frequencies of the target speaker and those of the reference speaker. This method has extended earlier work by a number of researchers on using single VT-length parameter to normalize speaker differences (e.g., [1, 2]). The proposed method also offers more detailed frequency warping functions than the conventional VT-length normalization method but requires only a small amount of adaptation data to estimate the two dimensional VT geometric parameters. To evaluate the proposed new speaker normalization method, we conducted continuous speech recognition experiments using speaker-normalized HMMs. The results of the recognition experiments in comparison with unnormalized gender-dependent HMM and with a number of conventional speaker-normalization methods demonstrate the effectiveness of the new method in terms of performance and the required adaptation data for obtaining the speaker-normalization parameters.

2. THE SPEAKER NORMALIZATION METHOD

In this method, we first approximate the VT model of the target speaker based on the following two factors: 1) phoneme-dependent factor: A standard area function that depends on phonemes; and 2) speaker dependent factor: Two dimensional VT geometric parameters including the length of the oral section (l_1) and the pharyngeal section (l_2) of the VT. Then speaker and phoneme-dependent VT shapes are approximated, and the acoustic features computed based on the approximated VT shapes are used for estimating frequency warping functions for speaker normalization. In this section, we will describe how the VT model of the target speaker is approximated and used for estimating frequency warping functions

2.1. Estimation of VT parameters

In this study, we use two VT geometric parameters to characterize each speaker: the length of the oral section of the VT (l_1) and the length of the pharyngeal section of the VT (l_2). The main motivation for using these more detailed VT geometric parameters (instead of single VT length, for example) comes from the well known observation of non-uniform formant scaling over a frequency range much greater than what can be accounted for by a single factor of VT-length variation [4]. In this work, the two VT parameters, l_1 and l_2, that jointly characterize the gross VT geometry of a speaker are estimated from formant frequencies (F1, F2, F3) of two Japanese vowels (/a/ and /i/) that reach their targets. The l_1 and l_2 values are fixed for a stylized reference speaker's VT. Given the information about VT constriction and the approximate area function for a particular vowel based on the stylized reference speaker's VT geometry, the formant frequencies of that vowel are computed for any new VT geometry associated with the target speaker that is characterized by the target speaker's estimated l_1 and l_2 values. The computation is carried out by artificially stretching (or shrinking) the reference speaker's l_1 and l_2 lengths separately in a linear fashion. Given a vowel, two independent stretch (or shrink) factors (for l_1 and l_2, respectively) are mapped to a set of formants according to the model computation. The formant space is generated by a chosen set of vowels and by a full range of stretch (or shrink) factors (limited to possible vowel phonetic-identity changes). To facilitate inverse mapping from formants to the stretch factors, the formant space is approximated by piece-wise linear functions built from a large number of points computed from the model. Each piece-wise linear function is confined within a corresponding triangle grid of points in the domain of stretch factors.

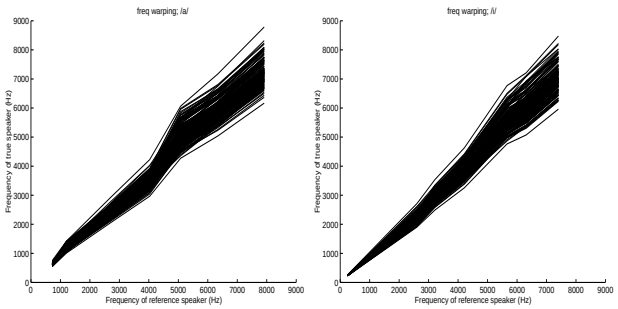
Once the mapping function between the formant space and the stretch factors is formed (all based on the VT model computation), then given the formant data (target vector) of vowels from any new speaker, a search is conducted to find the stretch factors whose mapped formant vector will be as close to the target formant vector as possible. The stretch factors thus found are multiplied by the l_1 and l_2 values of the reference speaker to give the l_1 and l_2 values for the new speaker.

2.2. Phoneme-dependent frequency warping function generation for speech recognition

The non-uniform formant scaling results described in [4] have clearly suggested phoneme dependence in inter-speaker frequency warping. Accordingly, in this method, phoneme-dependent VT shapes are estimated by manipulating the control parameters of the articulatory model and are used for estimating frequency warping function according to the following procedure: (1) construct a phoneme-dependent shape of the VT tubes of the reference speaker; (2) approximate the target speaker's phoneme-dependent VT shape by stretching or shrinking the VT tube of the reference speaker according to the l_1 and l_2 values of the target speaker; (3) compute the several lowest resonance frequencies (e.g., F1-F7) of the VT shapes of the reference speaker and target speaker, respectively. Then, the relationship between the F1-F7 values computed in step (3) above gives an estimation of the (nonlinear) frequency warping function from the target speaker to the reference speaker for the VT shape depending on the phoneme. In order to fill in the frequency points not covered by the F1-F7 values, linear interpolation is used for the frequency regions between each of the adjacent resonance-frequency values.

We have run the above frequency procedure for estimating warping function with a total of 148 male speakers.[1] The results for the two Japanese phonemes /a/ and /i/ are shown in Fig 1, where each curve corresponds to a separate target speaker.

[1] the same speakers used in speech recognition experiments described in Section 3.



(1) Vowel /a/ (2) Vowel /i/

Figure 1. Phoneme-dependent frequency warping functions for Japanese vowels and for 148 target speakers

For recognition, the same procedure with the conventional one-pass Viterbi algorithm is used for recognition with the following two differences. First, some separate feature parameters are calculated using the estimated phoneme-dependent frequency warping functions. Second, for each phoneme's HMM states, the separate feature parameters are used to calculate the HMM output probabilities. The training method of phoneme-dependent speaker normalized HMMs is also the conventional Viterbi training algorithm, except the output probabilities of each HMM state are calculated from the feature parameters analyzed by using phoneme-dependent frequency warping functions.

3. EXPERIMENTS

3.1. Conditions

This section reports our evaluation experiments on the proposed speaker normalization method using a Japanese 26-phoneme recognition task. The experimental conditions are listed in Table 1. The ATR phonetically balanced sentence speech database (Cset)[6] was used for the experiments. The following two types of test speaker sets were selected from the 148 speakers of this database: (1) 10 randomly selected speakers (**random**), (2) 10 speakers with the lowest accuracy of speech recognition experiments using gender-dependent HMMs trained with all 148 male speaker's speech data (**worst**). The remaining 128 speaker's data are used to train speaker-normalized HMMs. These HMMs are trained with 50 Japanese phonetically balanced sentences (a total of 2774 phonemes) uttered by the 128 male speakers.

Phoneme recognition experiments were performed using the one-pass Viterbi algorithm with the syllabic constraints of the Japanese language expressed as a phoneme-pair grammar. The test data consisted of 50 sentences (a total of 2905 phonemes) per speaker.

Acoustic Analysis	
Sampling frequency 12kHz, Hamming window 20ms Frame period 5ms, Filter bank order 16, log power + 12-th MFCC + Δ log power + 12-thΔMFCC	
Topology of HMM	
1000 states tied state context-dependent HMM[5] with 3 states 10 mixtures pause model	
Estimation data of VT parameters	
Two Japanese vowels /a/ and /i/ extracted from two words "y-u-u-z-a-a" and "f-a-m-i-r-i-i".	
Training data	
128 male speakers (50 sentences/person)	
Recognition data	
random :	10 male speakers (50 sentences/person) (randomly selected from 148 speakers)
worst :	10 male speakers (50 sentences/person) (worst 10 speakers of 148 speakers)

Table 1. Experimental Conditions

3.2. Results

Table 2 shows the phoneme error rate obtained by using several separate speaker-normalization methods. Speaker-normalized HMMs with 5 Gaussian mixtures were trained by using the following speaker-normalization methods: 1) gender-dependent model (**GD**); 2) vocal-tract length normalization (**VTLN**)[2]; and 3) phoneme-dependent speaker normalization (**L1L2**)[3] [4]. The left side of the table shows results of VT-based normalization only, and the right side shows results obtained by using VT-based normalization and cepstral mean normalization (CMN), which is well known to efficiently reduce distortion of spectral tilt.

The results show that VT-based speaker normalization methods have reduced phoneme recognition errors by about 10% from the GD model. The greatest error reduction comes from phoneme-dependent speaker normalization

[2] In the case of **VTLN**, frequency warping function is defined as $f' = \frac{\overline{VTL_{training}}}{VTL_{target}} \times f$, where VTL_{target} is vocal-tract length of target speaker, and $\overline{VTL_{training}}$ is the average of the vocal-tract lengths of the 128 training speakers.

[3] The proposed frequency warping functions and speaker normalization methods are strongly influenced by the acoustic property of the reference speaker. To reduce this factor, we defined frequency warping functions as resonance frequencies of target speaker warped into the average of resonance frequencies of 128 training speakers.

[4] Phoneme-dependent frequency warping functions for five Japanese vowels and one for another phonemes approximated based on neutral VT shapes (close to that corresponding to schwa) are prepared for normalization.

model	test-sets (MFCC)			test-sets (CMN)		
	random	worst	Ave.	random	worst	Ave.
GD	15.36	28.16	21.76	14.31	23.66	18.99
VTLN	14.45	24.88	19.66	14.52	20.58	17.55
L1L2	14.45	24.54	19.49	14.38	20.55	17.47

Table 2. Phoneme recognition error rates(%) obtained by using four speaker-normalization methods.

method **L1L2**, which reduces phoneme recognition error rates by 5.9% for test-set **random** and 13% for test-set **worst**.

Furthermore, the use of CMN reduces recognition error for test-set **worst** by 16% from the GD model without CMN, and the use of the proposed frequency warping based normalization with CMN reduced recognition error by 27% from the unnormalized GD model.

Furthermore, we have compared the proposed VT-based speaker normalization methods with the conventional data-driven speaker adaptation approach. We conducted the same phoneme recognition experiments using a speaker-adapted model trained by "Transfer Vector Field Smoothing Methods" (VFS) [7]. The phoneme recognition error rate for test-set **worst** is given in Table 3. In the case of VFS, the speaker-adapted model, trained by changing the amount of speech data for adaptation, was used for experiments. The results show that our new speaker-normalization method achieves a performance that is equivalent to that of the speaker-adapted model established by VFS using three sentences per speaker while requiring only two vowels per speaker to estimate the VT parameters for use in speaker normalization.

4. CONCLUSIONS

We have proposed a speaker normalization method for speech recognition using model-based vocal-tract (VT) geometric parameters. The nonlinear and phoneme-dependent frequency warping functions, which are estimated based on the formant frequencies of each speaker's crude VT geometry, are used to normalize the acoustic properties of speakers. The results in recognition experiments show that this method is superior in performance to the conventional data-driven speaker adaptation and normalization methods while drastically reducing the amount of adaptation data needed to estimate the speaker normalization parameters. On the other hand, the results of recognition experiments using the proposed speaker normalization with CMN shows

GD	VFS (adapted with N sentences)					L1L2
	1	2	3	6	10	
28.16	27.13	25.82	25.42	22.96	20.38	24.54

Table 3. Phoneme recognition error rates(%) obtained by using speaker normalized model and speaker adapted model with N sentences by VFS.

that the effectiveness of speaker normalization on frequency domain is not adequate. Therefore, we are planning to apply the acoustic feature estimated from the VT-model to normalize speaker using the power spectrum.

Acknowledgments:
We would like to thank Dr. A. Galvan for the Matlab codes used to estimate the VT parameters. We are also grateful to Dr. Yamamoto, President, ATR ITL Laboratories and all of the members of Dept. 1 for their advice and encouragement.

REFERENCES

[1] E. Eide and H. Gish: "A parametric approach to vocal tract length normalization," Proc. of ICASSP, 1996, pp. 346-349.

[2] P. Zhan and M. Westphal: "Speaker normalization based on frequency warping," Proc. of ICASSP, 1996, pp. 1039-1042.

[3] M. Naito, L. Deng and S. Sagisaka: "Speaker clustering for speech recognition using the parameters characterizing vocal-tract dimensions," Proc. of ICASSP, 1998, pp. 981-984.

[4] G. Fant: "Non-uniform vowel normalization," Speech Transmission Laboratory Quarterly Progress and Status Report, Vol.2-3, 1975, pp. 1-19.

[5] J. Takami and S. Sagayama: "A successive state splitting algorithm for efficient allophone modeling," Proc. of ICASSP, 1992, pp. 573-576.

[6] T. Takezawa, T. Morimoto and Y. Sagisaka: "Speech and Language Databases for Speech Translation Research in ATR," Proc. of EALREW, 1998, pp. 148-155.

[7] K. Ohkura, M. Sugiyama and S. Sagayama: "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs," Proc. of ICSLP, 1992, pp. 369-372.