SPEAKER CLUSTERING FOR SPEECH RECOGNITION USING THE PARAMETERS CHARACTERIZING VOCAL-TRACT DIMENSIONS

Masaki Naito[1], *Li Deng*[1,2], *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 propose speaker clustering methods based on the vocal-tract-size related articulatory parameters associated with individual speakers. Two parameters characterizing gross vocal-tract dimensions are first derived from formants of speaker-specific Japanese vowels, and are then used to cluster a total of 148 male Japanese speakers. The resultant speaker clusters are found to be significantly different from the speaker clusters obtained by conventional acoustic criteria. Japanese phoneme recognition experiments are carried out using speaker-clustered tied-state HMMs(HMNets) trained for each cluster. Compared with the baseline gender dependent model, 5.7% of recognition error reduction has been achieved based on the clustering method using vocal-tract parameters.

1. INTRODUCTION

Use of speaker-clustered models is a simple but effective way to improve the accuracy of speaker-independent speech recognition, which has been clearly exemplified by use of Gender-Dependent (GD) models. However, within each gender, there is still a wide variety of speakers. To obtain more detailed speaker clusters, several researchers have proposed several methods. Kosaka and Sagayama, for example, proposed a tree-structured speaker clustering algorithm and a fast speaker adaptation method based on selection of appropriate speaker clusters defined on the tree [5]. The effectiveness of this method was reported as an initialization model for speaker adaptation in [8]. Nearly all the previous work on speaker clustering was based on similarities across speakers defined by acoustic distances. In particular, the acoustic distances across speakers were quantified according to the speaker-dependent models used for speech recognition (e.g. HMMs).

We believe that characterization of speaker differences should be more effective using articulatory parameters than acoustic ones, and that this should be especially so within a gender group where the acoustic differences across genders have been drastically reduced. One reason, among several others, for this is that the vocal-tract (VT) geometric differences across speakers, which account for a large portion of the overall speaker differences, can be easily and naturally characterized by intuitions using low-dimensional parameters. On the other hand, the acoustic differences, which reflect (although not all) the VT geometric differences in a highly nonlinear fashion, must be characterized by high-dimensional parameters not easily subject to physical interpretation but easily giving rise to local optimum during cluster training. This consideration forms the motivation of the work reported in this paper, where VT parameters related to gross VT dimensions are used to cluster a total of 148 male Japanese speakers in our database. Two clustering methods, with and without a tree structure, are implemented using either acoustic parameters (baseline) and the VT parameters. The clustering methods have been evaluated in Japanese phoneme recognition experiments using speaker clustering tied-state HMMs (SC-HMNet)[7]. The results show that the performance of SC-HMNets based on VT parameters is higher than those of GD-HMNet and of the SC-HMNets for clusters based on acoustic parameters.

Use of the VT parameters as reported in this paper will offers a way of quick adaptation since potentially two vowel tokens are sufficient to estimate these parameters and to select the most appropriate speaker cluster. There is no need to use large data as are required for acoustics-based schemes. This work represents our initial effort in pursuing production-based modeling for speech recognition, and can be seen as a simple extension of previous works on use of one-dimensional VT-length (e.g. [2]) to two-dimensional VT parameters. Although the gain obtained so far has not as striking as expected, it is promising enough to warrant further extension of this work to more sophisticated speaker adaptation schemes.

2. SPEAKER CLUSTERING METHODS

In this section, we describe two types of speaker clustering methods (together with the distance measures) used in this work, one with use of a tree structure, the other with use of a flat, plain structure in organizing the clustered speakers. Both methods have been used for acoustic and VT parameters.

2.1. Plain speaker-clustering algorithm

In this clustering method, all the distances (Bhattacharyya or Euclidean; see details later) between speakers are calculated in advance and a distance table is created. The cluster with the maximum sum of distances is divided using the distance table[6]. In this algorithm, the fixed number of clusters or a distance threshold value is required to stop the flat-structured cluster splitting and growing.

2.2. Tree-structured speaker clustering algorithm

In the tree-structured clustering algorithm, a fixed number K controls the number of sub-clusters at each node. This procedure enables all speakers to be hierarchically clustered.

Details of the algorithm are:

STEP 1 Set $j = 1$. All speakers are clustered by the plain clustering method, and then K centroid speakers $\{m_1(j), \ldots, m_K(j)\}$ are obtained (j denotes the hierarchical level of the tree).

STEP 2 If the number of speakers satisfying $s \in S_l(j)$ becomes fewer than K, quit clustering for cluster l.

STEP 3 For the l-th cluster $S_l(j)$, except those that quitted in the previous step, the speakers satisfying $s \in S_l(j)$ are clustered to produce K sub-clusters. This creates the next-level, new K speaker clusters $M^l(j+1) = \{m_1^l(j+1), \ldots, m_K^l(j+1)\}$.

STEP 4 $j \leftarrow j + 1$. Return to STEP 2.

2.3. Distance measures

We use two distance measures between speakers: one suitable for acoustic parameters (Bhattacharyya), the other suitable for VT parameters (Euclidean). For the first case, speaker-dependent HMNets (SD-HMNets) of an identical structure are trained first by the Baum-Welch algorithm. Then the distance between two speakers is defined as average of the Bhattacharyya distance between output probability functions of each speaker's SD-HMNet [5]; that is, for two different SD-HMNets, M_1 and M_2, the distance is

$$D(B^{(1)}, B^{(2)}) = \frac{1}{MN} \sum_{i,j=1}^{N} \sum_{k=1}^{M} d(b_{ij}^{(1)}(k), b_{ij}^{(2)}(k)), \qquad (1)$$

where

$$d(b^{(1)}, b^{(2)}) = \frac{1}{8}(\mu_1 - \mu_2)^t \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} (\mu_1 - \mu_2) + \frac{1}{2} \ln \frac{|(\Sigma_1 + \Sigma_2)/2|}{|\Sigma_1|^{1/2} |\Sigma_2|^{1/2}}, \qquad (2)$$

N is the total number of the HMNet states, and M the total number of the output distributions. Estimation of the VT parameters is performed by a functional mapping method described in Section3. The distance between two speakers used in clustering methods is defined as the Euclidean distance between the two speaker-specific low-dimensional vectors consisting of the VT parameters.

2.4. Speaker cluster selection

After speaker clustering is accomplished, SC-HMNets are trained using the Baum-Welch algorithm. Fast speaker adaptation is then performed by selecting a most suitable SC-HMNet for the target speaker.

Speaker cluster selection is based on the maximum likelihood criterion. The likelihood of a SC-HMnet is calculated according to a Viterbi procedure that uses short speech acoustic and transcription data (adaptation information) of each speaker. In the case of plain clustering model, the SC-HMnet that gives the maximum likelihood score is selected for the target speaker. In the case of tree-structured clustering model, the likelihood of each SC-HMNet is calculated at each level of the tree structure. The tree is then traced by selecting the SC-HMnet that gives the maximum likelihood score. At each level of the tree, the likelihood of the selected SC-HMNet is memorized. After the tree tracing, the SC-HMnet that gives the overall maximal likelihood score is selected by comparing the memorized likelihood at each tree level.

3. ESTIMATION OF VT PARAMETERS

The VT parameters used in this study for speaker clustering are of two types: 1) the length of the oral section of the VT (l_1) together with the length of the pharyngeal section of the VT (l_2) (two parameters); and 2) the total VT length (VTL; single parameter) which is sum of l_1 and l_2 ($VTL = l_1 + l_2$). The reason why we need to have more than a single VT length to characterize the VT comes from the clear evidence 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 [3]. The reasons why we choose l_1 and l_2 parameters in this study are: 1) two dimensions are a most straightforward extension from earlier one-dimension VT-length normalization work conducted by many research groups (cf. [2]); 2) the ratio of oral and pharyngeal section lengths of the VT is a significant factor shaping acoustic outputs of speech; this is so because phonetically significant VT constrictions are usually made with reference to the oral-section length of the VT (l_1) but the formants depend on the entire VT length including pharyngeal length l_2; and 3) methods for estimating l_1 and l_2 are relatively easy. In this section, we describe how Japanese vowel formant data from our evaluation database [1] are used to estimate the VT parameters l_1 and l_2 , which are then used to determine the Euclidean distance between any pair of speaker-specific two-dimensional vectors of $\{l_1, l_2\}$ (and between any pair of speaker-specific scalars of VTL).

The estimation method for l_1 and l_2 parameters is based on an articulatory model developed previously for speaker normalization purposes (cf. [4]). The model characterizes the gross VT geometry of a speaker (which is independent of phonetic units) by two parameters: the length of the oral section (l_1) and the length of the pharyngeal section (l_2). The l_1 and l_2 values are fixed for a stylized reference speaker's VT. Given information about VT constriction and a related approximate area function for a particular vowel with the stylized reference speaker's VT geometry, the model is capable of computing the formant frequencies of that vowel for any new VT geometry obtained artificially by independent linear stretch (or shrink) of the reference speaker's l_1 and l_2 lengths. [2]

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 by possible vowel phonetic-identity changes). To facilitate inverse mapping from formants to the stretch factors, the formant space is approximated by piecewise linear functions built from a large number of points computed from the model. Each piecewise linear function is confined within a corresponding triangle grid of points in the domain of stretch factors.

[1] In our work, formant frequencies (F1,F2,F3) of two Japanese vowels /a/ and /i/ are obtained for each speaker. Each vowel is extracted from two words of speech database uttered phrase-by-phrase. The vowel /a/ is extracted from Japanese word "b-a-a-i", and /i/ from "f-a-m-i-r-i-i".

[2] The limit of the linear stretch is 130%, and that of the linear shrink is 70%. Beyond these limits, some vowels will change their phonetic identities (according to informal listening of the synthesized vowels) after the stretch or shrink.

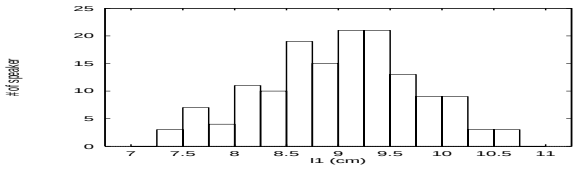

Figure 1: Distribution of the VT parameter l_1

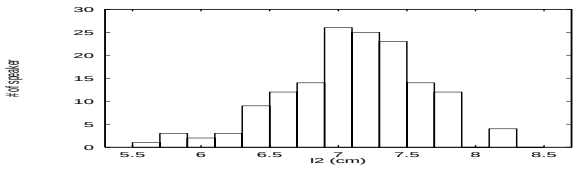

Figure 2: Distribution of the VT parameter l_2

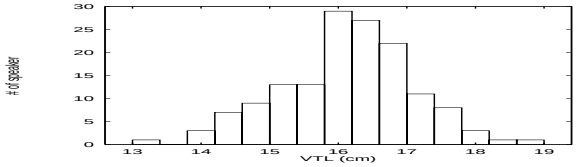

Figure 3: Distribution of the VT length (VTL)

Once the mapping function between the formant space and the stretch factors is formed (all based on the articulatory 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.

4. RESULTS OF SPEAKER CLUSTERING

A total of 148 male speakers were clustered based on both the acoustic data and on the VT parameters. Before we show clustering results, we first show the distributions (over all 148 male speakers) of the estimated VT parameters, including l_1 and l_2, as well as their sum $VTL = l_1 + l_2$, in Figs.(1),(2), and (3), respectively. The means of these parameters over the speakers are l_1 =9.01 cm, l_2 =7.10 cm, and VTL =16.11cm, respectively. We note that the distributions are fairly smooth over the VT parameters, with no signs of bimodal distributions. This is consistent with earlier results on English speech for gender-specific VTL-related parameters (cf. [9] for frequency warping factors). Properties of l_1 and l_2 distributions have not been studied in the past, and it is interesting to observe also the smooth distributions illustrated in Figs.(1) and (2).

To calculate the acoustic distance between speakers, 200-state unimodal Gaussian SD-HMNet is trained. Each SD-HMNet is trained individually (i.e. speaker by speaker) with 50 common Japanese phonetically balanced sentences. The Baum-Welch algorithm with controlled variance is then used for training each SD-HMNet.

In the case of plain clustering, all 148 male speakers are clustered into 3, 5, 10, 20, or 40 clusters. In the case of tree structured clustering, these speakers are clustered into five clusters at each node of the clustering tree. Plain-clustering results for the five-cluster case are shown in Fig.(4), (5), and (6), respectively, based on the estimated VTL information (Euclidean distance), the estimated l_1 plus l_2 information (Euclidean distance), and acoustic information (Bhattacharyya distance). Each point in these figures represents a distinct speaker specified by his l_1 and l_2 dimensions. All the speakers belonging to the same cluster are represented by the same symbol.

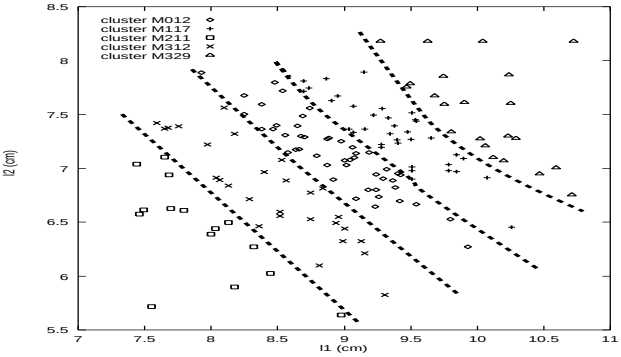

Figure 4: Clusters obtained using VTL

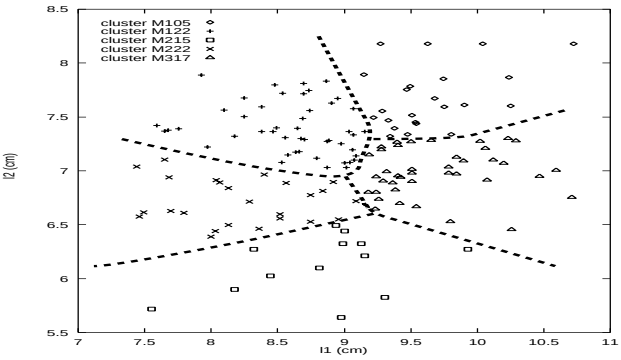

Figure 5: Clusters obtained using l_1 plus l_2

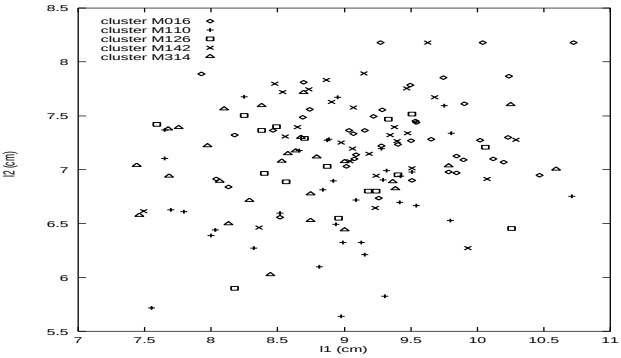

Figure 6: Clusters obtained using acoustic data

From the results shown in Figs.(4), (5), and (6), we observe drastically different clusters using acoustic and vocal-tract parameters. Fig.(6) demonstrates that the acoustically clustered speaker groups do not correlate with the geometrical differences of the speakers. On the other hand, the clusters obtained from the VTL information (Fig. 4) and those from the l_1 plus l_2 information are highly related to each other (Fig.5). The latter is expected since one set of information is derived from the other, and the consistency shown here verifies correct implementation of the clustering procedure.

Analysis	Sampling frequency 12kHz
	Hamming window 20ms
	Frame period 5ms
	16-th LPC Cepstrum
	+16-thΔLPC Cepstrum+Δlog power
Training data	148 males (50 sentences per person)
Recognition data	6 males
· Cluster Selection	7 phrases per person (SB1 task)
· Recognition	249 phrases (SB1 task)

Table 1: Experimental Conditions

5. SPEECH RECOGNITION EXPERIMENTS

5.1. Experimental conditions and data sets

In this section, we report our evaluation experiments on the various speaker clustering methods described in this paper on a Japanese 26-phone recognition task. The experimental conditions are listed in Table 1. Given the clusters determined as described in earlier sections, each SC-HMNet (containing 200 states of unimodal Gaussian) is trained using the Baum-Welch algorithm (with variance controlled) with 50 Japanese phonetically balanced sentences (a total of 2774 phones) uttered by all 148 male speakers. The GD-HMNet (i.e., single-cluster HMNet) is trained with the same data. Speech data consisting of seven phrases (containing 51 phones) are used to select speaker cluster. The test data consist of 249 phrases (a total of 1963 phones) in phoneme recognition experiments.

5.2. Recognition results

Table 2 presents the comparative results of phoneme recognition accuracy obtained by using the SC-models. Horizontally arranged performance numbers are associated with the following speaker-cluster conditions: 1) Gender Dependent model **(GD)**; 2) 3, 5, 10, 20, and 40 speaker clusters by plain clustering algorithm; and 3) tree-structured speaker clustering. Vertically arranged performance numbers denote the following information used for clustering: 1) acoustics (**Acoust**); 2) vocal tract length (**VTL**); and 3) vocal tract length of oral section and pharyngeal section (l_1/l_2).

These results demonstrate that use of the SC models reduces phoneme recognition errors by 0.2-5.7% compared with the GD model. The greatest error reduction (5.7%) comes from the SC-HMNets trained for five plain speaker clusters based on two-dimensional VT parameters l_1 plus l_2, In general, use of l_1 plus l_2 parameters gives the highest performance, followed by use of VTL parameter. Use of acoustic information gives the least amount of performance improvement.

6. DISCUSSIONS AND SUMMARY

Earlier results have shown the effectiveness in speech recognition of using general articulatory parameters to provide a natural means of modeling contextual variations of speech [1]. This work shows how the articulatory parameters which specify gross VT dimensions can be used to naturally and economically represent speaker variations in the speech. The specific scheme used in this work is to cluster speaker groups according to their VT-dimension parameters. Variabilities in these parameters reflect one significant physical cause accounting for the observed acoustic differences in the speech signal which is generated from the VT.

We have proposed a speaker clustering method using the VT parameters. In this method, the VT parameters are estimated from formants of only two Japanese vowels based on functional mapping from the formant space to the VT parameter space. Both plain and tree-structured speaker clusterings are created based on the estimated VT parameters. The results of speaker clustering show that there is little correlation between the obtained clusters based on acoustic data and those on VT parameters.

Methods	GD	plain clustering (# of cluster)					tree
		3	5	10	20	40	
Acoust.	66.5	67.9	67.0	66.6	66.9	66.2	67.2
VTL	66.5	67.7	67.5	68.0	67.2	66.7	68.2
l_1/l_2	66.5	67.7	68.4	68.3	68.0	67.5	67.2

Table 2: Recognition results using SC-HMNets (%)

The effectiveness of our speaker clustering method has been shown in Japanese phoneme recognition experiments using the SC-HMNets. Compared with the baseline GD model, 5.7% recognition error reduction is obtained by using the SC-HMNets which are trained for the clusters constructed based on the VT parameters. This performance is also higher than that of the SC-HMNets obtained by clustering based on the acoustic distance measure.

We are planning to expand the number of speakers used for clustering (including female speakers), and investigate other parameters specifying VT shapes rather than those specifying only the gross VT geometry as reported in this paper. Further, use of the VT parameters for speaker normalization and adaptation will be investigated.

Acknowledgments: We would like to thank Dr. A. Galvan who originally developed and contributed the Matlab codes for estimating the vocal tract parameters. We are also grateful to Dr. Yamamoto, President, ATR ITL Laboratories and all of the members of Dept. 1 for their advices and encouragements.

7. REFERENCES

[1] L. Deng and D. Sun: "A statistical approach to ASR using atomic units constructed from overlapping articulatory features," *J. Acoust. Soc. Am.*, Vol.95, 1994, pp. 2702-2719.

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

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

[4] A. Galvan and L. Deng. "Speaker-independent phonetic classification and recognition using an articulatory model for formant-space speaker normalization," in preparation.

[5] T. Kosaka and S. Sagayama: "Tree-Structured Speaker Clustering For Fast Speaker Adaptation," Proc. of ICASSP, 1994, pp. 245-248.

[6] N. Sugamura, K. Shikano and S. Furui: "Isolated Word Recognition Using Phoneme-Like Templates," Proc. of ICASSP, 1983, pp. 243-252.

[7] J. Takami and S. Sagayama: "A Successive State Splitting Algorithm for Efficient Allophone Modeling," Proc. of ICASSP, 1992, pp. 573-576.

[8] M. Tonomura, T. Kosaka and S. Matsunaga: "Speaker Adaptation Based on Transfer Vector Field Smoothing Using Maximum a Posteriori Probability Estimation," Proc. of ICASSP, 1995, pp. 688-691.

[9] P. Zhan and M. Westphal: "Speaker Normalization Based on Frequency Warping," Proc. of ICASSP, 1996, pp. 1039-1042.