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

Masaki Naito¹, Li Deng¹,², Yoshinori Sagisaka¹

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 (SC-HMNs) 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 acoustics, 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-HMNs) [7]. The results show that the performance of SC-HMNs based on VT parameters is higher than those of GD-HMNets and of the SC-HMNs for clusters based on acoustic parameters.

Use of the VT parameters as reported in this paper will offer 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 acoustic-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 $s \in S(j)$
becomes fewer than $K$, quit clustering for cluster $l$.

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

STEP 4 $j = j + 1$. 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 HM Nets (SD-HM Nets) of an identical
structure are trained first by the Baum-Welch algorithm.
Then the distance between two speakers is defined as an
average of the Bhattacharyya distance between output prob-
ability functions of each speaker’s SD-HM Net [5]; that is, for
two different SD-HM Nets, $M_1$ and $M_2$, the distance is

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

where

$$d(b^{1}, b^{2}) = \frac{1}{8} \left( \mu_1 - \mu_2 \right)^{2} \left( \frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} \left( \frac{\mu_1 - \mu_2}{4} + \frac{1}{\Sigma_1 + \Sigma_2} \right)$$

$$+ \frac{1}{2} \ln \left( \frac{\Sigma_1 + \Sigma_2}{2} \right)^{1/2}. \quad (2)$$

$N$ 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 map-
ing method described in Section 3. 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-HM Nets
are trained using the Baum-Welch algorithm. Fast speaker
adaptation is then performed by selecting a most suitable
SC-HM Net for the target speaker.

Speaker cluster selection is based on the maximum like-
lihood criterion. The likelihood of a SC-HM Net 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-HM Net that gives the maximum likelihood score is se-
lected for the target speaker. In the case of tree-structured
clustering model, the likelihood of each SC-HM Net is cal-
culated at each level of the tree structure. The tree is then
traced by selecting the SC-HM Net that gives the maximum
likelihood score. At each level of the tree, the likelihood
of the selected SC-HM Net is memorized. After the tree trac-
ing, the SC-HM Net 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 articular 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 refer-
ence speaker’s VT. Given information about VT constric-
tion and a related approximate area function for a particu-
lar 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

Following a vowel, two independent stretch (or shrink) fac-
tors (for $l_1$ and $l_2$, respectively) are mapped to a set of
formants according to the model computation. The form-
ment 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 corre-
sponding 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 “tsa-a”,
and /i/ from “tsa-i-mii”.

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.
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 \( l_1 = 9.01 \text{ cm}, \ l_2 = 7.10 \text{ cm}, \) and \( VTL = 16.11 \text{ cm} \), 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 VT-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, a 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-Welsh 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 VT 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.

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 VT 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.
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 48 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 1063 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 (l1/l2).

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-HMNet trained for five plain speaker clusters based on two-dimensional VT parameters l1 plus l2. In general, use of l1 plus l2 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.

The effectiveness of our speaker clustering method has been shown in Japanese phoneme recognition experiments using the SC-HMNet. Compared with the baseline GD model, 5.7% recognition error reduction is obtained by using the SC-HMNet which are trained for the clusters constructed based on the VT parameters. This performance is also higher than that of the SC-HMNet 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, ATRITL Laboratories and all of the members of Dept. 1 for their advice and encouragements.

7. REFERENCES