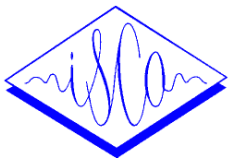

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

SPEAKER-CONSISTENT PARSING FOR SPEAKER-INDEPENDENT CONTINUOUS SPEECH RECOGNITION

Kouichi Yamaguchi[†1], *Harald Singer*[†2], *Shoichi Matsunaga*[†2] *and Shigeki Sagayama*[†3]

[†1]SHARP Corporation, Information Technology Research Labs., Tenri-shi, Nara 632, Japan
[†2]ATR Interpreting Telecommunications Research Labs., Seika-cho, Kyoto 619-02, Japan
[†3]NTT Human Interface Labs., Yokosuka-shi, Kanagawa 238-03, Japan

ABSTRACT

This paper describes a novel speaker-independent speech recognition method, called "speaker-consistent parsing", which is based on an intra-speaker correlation called the speaker-consistency principle. We focus on the fact that a sentence or a string of words is uttered by an individual speaker even in a speaker-independent task. Thus, the proposed method searches through speaker variations in addition to the contents of utterances. As a result of the recognition process, an appropriate standard speaker is selected for speaker adaptation.

This new method is experimentally compared with a conventional speaker-independent speech recognition method. Since the speaker-consistency principle best demonstrates its effect with a large number of training and test speakers, a small-scale experiment may not fully exploit this principle. Nevertheless, even the results of our small-scale experiment show that the new method significantly outperforms the conventional method. In addition, this framework's speaker selection mechanism can drastically reduce the likelihood map computation.

1 INTRODUCTION

A speech recognition task can take one of two approaches. One approach uses speaker-independent acoustic models, which has been used to develop many systems [1, 2]. This approach allows speakers to use the recognition system without any training procedure, but it provides no means of improving recognition performance for speakers who cannot be easily recognized. The other approach adapts acoustic models to the input speaker by using a speaker adaptation method such as the Vector Field Smoothing (VFS) [3]. It can deal with speakers who are hard to recognize by using adaptation methods.

Conventional recognition approaches, on the other hand, do not take into account the fundamental constraint of human speech production. Therefore, they risk smearing the speaker distribution with their mixture densities, which do not carry any assumptions of intra-speaker correlation. A sentence or a string of words is uttered by an individual speaker even in a speaker-independent task. We call this intra-speaker correlation the "speaker-consistency principle." Niyogi [4] exploited this intra-speaker correlation to improve the vowel recognition performance. Imamura [5] proposed a stochastic speaker classifier which determines the feature subspace suitable for an input speaker. In the acoustic HMMs, the observation emission probabilities are presented as joint probabilities for speaker individuality obtained from the speaker classifier and feature vectors from the acoustic pre-processor. In addition, it was found that gender-specific approaches significantly improve speaker-independent continuous speech recognition [6].

We introduce a novel speaker-independent speech recognition method, called "speaker-consistent parsing" ("speaker parsing" for short), which is based on the speaker-consistency principle. This method searches through speaker variations in addition to the contents of utterances. As a result of this recognition process, an appropriate standard speaker is selected for succeeding speaker adaptation. This recognition method can thus be utilized for speaker adaptation. This new method is experimentally compared with a conventional speaker-independent speech recognition method. Results show that this framework outperforms the conventional method and that it has the potential to efficiently adapt acoustic models to the input speaker by using a speaker selection mechanism.

2 SPEAKER PARSING PARADIGM

2.1 Mathematical Formulation

In this section we will provide the mathematical formulation for the speaker consistency principle.

Let $\boldsymbol{w}$ be a string of words $\boldsymbol{w} = w_1, w_2, \cdots, w_n$. Given acoustic evidence observation $\boldsymbol{y}$, the operations of speech recognition are to find the most likely word string, $\hat{\boldsymbol{w}}$, satisfying

$$P(\hat{\boldsymbol{w}}|\boldsymbol{y}) = \max_{\boldsymbol{w}} P(\boldsymbol{w}|\boldsymbol{y}). \quad (1)$$

The right-hand side of the above equation can be rewritten according to Bayes' rule as

$$P(\boldsymbol{w}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{w})P(\boldsymbol{w})}{P(\boldsymbol{y})}. \quad (2)$$

It follows from equations (1) and (2) that the purpose of the recognition operation is to find the word string $\hat{\boldsymbol{w}}$ that maximizes the product

$$P(\hat{\boldsymbol{w}})P(\boldsymbol{y}|\hat{\boldsymbol{w}}) = \max_{\boldsymbol{w}} P(\boldsymbol{w})P(\boldsymbol{y}|\boldsymbol{w}), \quad (3)$$

[1]Most of this work was done when K. Yamaguchi was a researcher at ATR Interpreting Telecommunications Res. Labs.

where $P(\boldsymbol{y}|\boldsymbol{w})$ relies on acoustic pattern matching.

Based on this, we apply the speaker-consistency principle to equation (1) as follows:

$$P(\hat{\boldsymbol{w}}, \hat{\boldsymbol{s}}|\boldsymbol{y}) = \max_{\boldsymbol{w},\boldsymbol{s}} P(\boldsymbol{w}, \boldsymbol{s}|\boldsymbol{y}), \quad (4)$$

where $\boldsymbol{s}$ is the speaker within m (*i.e.* $\boldsymbol{s} \in \{s_i|i = 1, 2, \cdots, m\}$). The right-hand side of the above equation can be rewritten according to Bayes' rule as

$$P(\boldsymbol{w}, \boldsymbol{s}|\boldsymbol{y}) = \frac{P(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{s})P(\boldsymbol{w})P(\boldsymbol{s})}{P(\boldsymbol{y})}, \quad (5)$$

where $P(\boldsymbol{s})$ is the *a priori* probability that any given test speaker is $\boldsymbol{s}$. In a speaker-independent speech recognition task, $P(\boldsymbol{s})$ can be thought of as equiprobable for all possible speakers. It follows from equations (4) and (5) that the purpose of the speaker parsing operation is to find the word string $\hat{\boldsymbol{w}}$ and the speaker $\hat{\boldsymbol{s}}$ that maximize the product

$$P(\hat{\boldsymbol{w}})P(\hat{\boldsymbol{s}})P(\boldsymbol{y}|\hat{\boldsymbol{w}}, \hat{\boldsymbol{s}}) = \max_{\boldsymbol{w},\boldsymbol{s}} P(\boldsymbol{w})P(\boldsymbol{s})P(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{s}), \quad (6)$$

where $P(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{s})$ is the probability that when the speaker $\boldsymbol{s}$ says the word string $\boldsymbol{w}$ the acoustic evidence $\boldsymbol{y}$ will be observed. Thus, this method searches through speaker variations $\{\boldsymbol{s}\}$ in addition to the contents of utterances $\{\boldsymbol{w}\}$. This recognition process selects an appropriate standard speaker $\hat{\boldsymbol{s}}$ that can be utilized for succeeding speaker adaptation.

Another formulation of the speaker-consistency principle is possible:

$$P(\hat{\boldsymbol{w}})P(\boldsymbol{y}|\hat{\boldsymbol{w}}) = \max_{\boldsymbol{w}} P(\boldsymbol{w}) \sum_{\boldsymbol{s}} P(\boldsymbol{s})P(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{s}). \quad (7)$$

In this case, the speaker $\hat{\boldsymbol{s}}$ has to be sought separately.

2.2 Application to Speaker Adaptation

The system starts from a speaker-independent mode. The speaker parsing process then selects a speaker $\hat{\boldsymbol{s}}$. Since this speaker $\hat{\boldsymbol{s}}$ is the closest speaker to the input speaker, it could be suitable as a standard speaker for speaker adaptation. For the speaker adaptation to be robust, the system may select more than one standard speaker. If a recognition result $\hat{\boldsymbol{w}}$ is used as a training signal, the speaker parsing method can be applied to an unsupervised speaker adaptation.

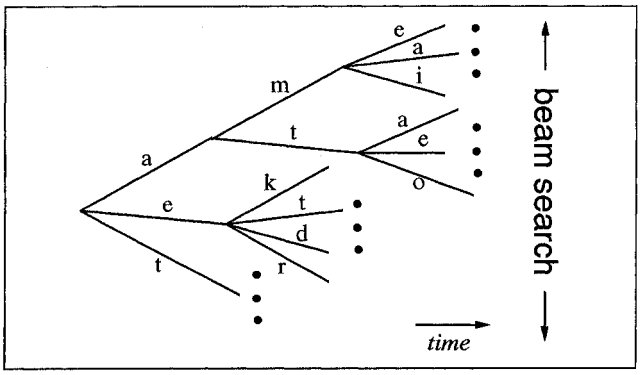

Figure 1. Parse tree (speaker-independent mode)

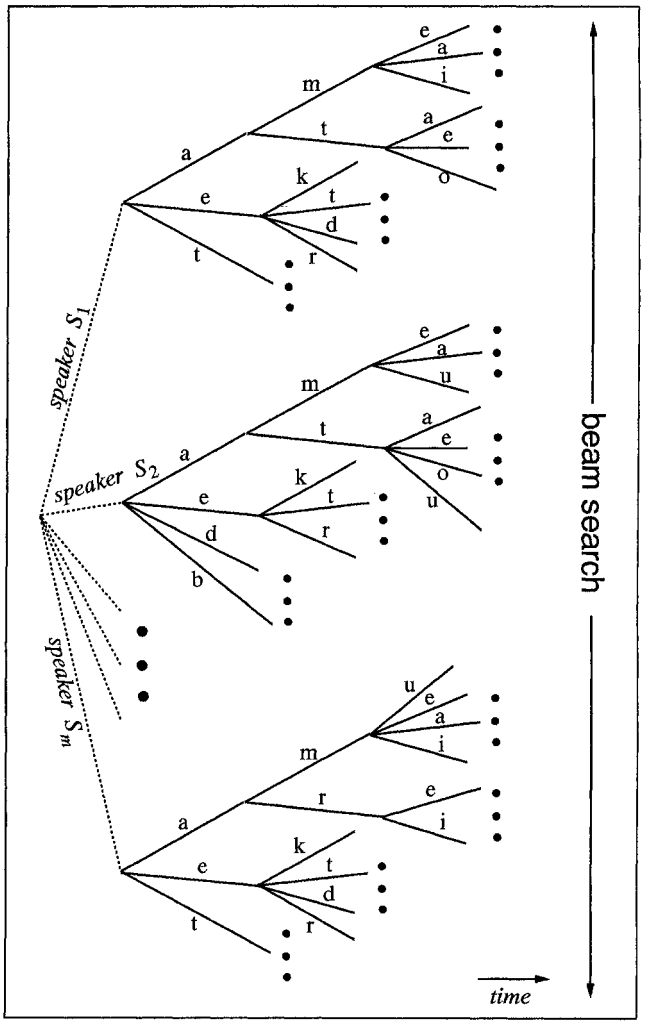

Figure 2. Parse tree (speaker parsing mode)

3 IMPLEMENTATION

3.1 Recognition System Overview

The speaker parsing method has been applied to the SSS-LR continuous speech recognition system [7]. This uses a hidden Markov network (HMnet) [8], which is an efficient representation of phoneme context-dependent HMMs. The HMnet is a highly generalized form of the HMM and incorporates context-dependent variations of phones and state sharing among different allophones. The system consists of a predictive LR parser and HMM allophone verifiers. The predictive LR parser is based on the generalized LR parser and used for phone prediction. The parser is guided by an LR table precompiled from context-free grammar rules.

3.2 Speaker Parsing Strategy

The SSS-LR system's beam search technique uses a phonetic parse tree. The phonetic parse tree in Figure 1 is illustrated for the conventional speaker-independent recognition mode. In this figure, the letter along each branch represents a phoneme predicted by the LR parser and verified by the allophone verifier.

The phonetic parse tree in Figure 2 is adopted for the speaker parsing recognition mode. We adopt equation (6) instead of equation (7) because the beam search equation (7) could only be used approximately. $\{P(\boldsymbol{s})\}$ is set equiprobable. At first there exist m hypotheses corresponding to m speakers $\{s_i | i = 1, 2, \cdots, m\}$. Then, the hypotheses grow separately, speaker by speaker; they are eventually pruned by the beam search technique as the tree growing process proceeds. From this point of view, this speaker parsing is different from the method that individually carries out m speech recognition processes with reference speakers $\{\boldsymbol{s}\}$ and selects the highest hypothesis at the end of the input utterance.

The HMnet for the speaker parsing is converted from a speaker-independent HMnet created by using a speaker-mixture algorithm [9]. Each HMM output probability density function is characterized by a 34-dimensional diagonal-covariance Gaussian mixture, and each Gaussian distribution is associated to a speaker $\boldsymbol{s}$. The speaker parsing method utilizes this speaker information to distinguish the hypotheses of different speakers.

4 EVALUATION (I)

The speaker parsing method is experimentally compared phrase by phrase with a conventional speaker-independent speech recognition method.

4.1 Experimental Conditions

Speaker-independent phrase speech recognition experiments were carried out using 18 test speakers (nine male and nine female). The test data involved 345 Japanese short phrases, *i.e.* the 1,600-word *International Conference Registration* task. Accordingly, the total number of test samples was 6,210 phrases. The context-free grammar included 2,813 rules with a phonetic perplexity of 3.3 and a morpheme perplexity of 53.8. The beam width was set at 1,200, where the recognition performance was almost saturated. The training data involved 310 Japanese short phrases different from the test data. The number of training speakers was 285 (142 male and 143 female).

The speaker-independent m-mixture HMnet was created by using a speaker-mixture algorithm as follows:

1. Generate a 200-state HMnet structure with an SSS algorithm.
2. Train 285 speaker-dependent single Gaussian HMnets using 310 phrases per speaker by VFS.
3. Select m HMnets out of 285 speaker-dependent HMnets by a tree-structured speaker clustering technique [10]. $(m = 2, 6, 10, 20)$
4. Retrain m HMnets using the samples of their cluster's members by VFS.
5. Merge m HMnets into an m-mixture HMnet.

4.2 Phrase Speech Recognition

The recognition results using several kinds of HMnets are tabulated in Table 1. Since the speaker-consistency principle best demonstrates its effect with a large number of training and test speakers, a small-scale experiment may not fully exploit this principle. Nevertheless, even in our small-scale experiment the speaker parsing method outperformed the conventional speaker-independent method for all number of mixtures. When the 6-mixture HMnet was used, the recognition rates of both methods were highest, and the difference between the top 1 recognition rates of both methods was 1.1%. This 6-mixture HMnet was used in the next section.

5 EVALUATION (II)

The previous experiments were carried out phrase by phrase. However, the comparison with the speaker-independent method should properly be carried out by using the speaker parsing method combined with a speaker adaptation method.

5.1 Experimental Conditions

The conditions of analysis, training data, test data and HMnets were the same as those of Evaluation (I). The 6-mixture HMnet was used here. The first 5, 10, 16 and 24 phrases of the test data in Evaluation (I) were selected for the training data of speaker adaptation. The test data were the final 340, 335, 329 and 321 phrases. The speaker adaptation method was a supervised training by VFS.

Two types of initial models for speaker adaptation were experimentally compared. One was a speaker-independent 6-mixture HMnet. The other was a single Gaussian HMnet which was extracted from the 6-mixture HMnet by a speaker selection mechanism of the speaker parsing method. The most selected speaker cluster $\boldsymbol{s}$ (*i.e.* single Gaussian HMnet) was decided as the standard speaker $\hat{\boldsymbol{s}}$ through all the speaker adaptation data by the speaker parsing method. This strategy of the speaker parsing method combined with a speaker adaptation is illustrated in Figure 3. $(n = 5, 10, 16, 24; N = 345)$

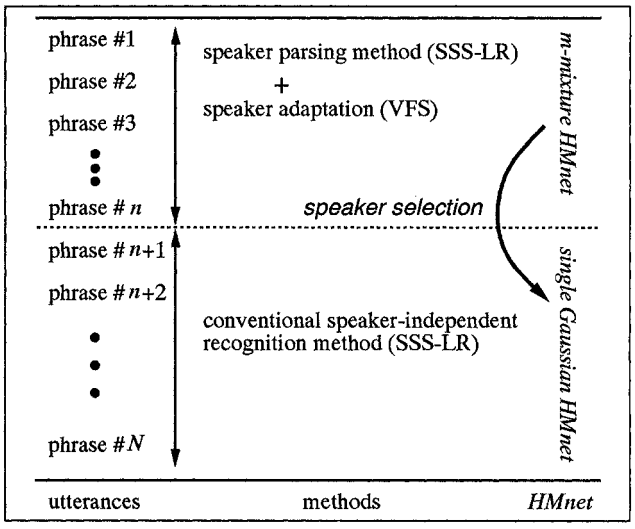

Figure 3. speaker parsing/adaptation strategy

5.2 Speaker Adaptation Experiments

The recognition experiment results using the two types of initial models for speaker adaptation are tabulated in Table 2. Although the single Gaussian HMnet had one-sixth degree of freedom, it outperformed the 6-mixture

Table 1. Speaker-independent phrase recognition rates (%)

recognition method	speaker parsing				speaker-independent			
# mixture (=cluster)	2	6	10	20	2	6	10	20
top 1	82.3	84.3	83.9	83.8	80.1	83.2	82.6	82.6
top 3	93.0	94.0	93.9	93.8	91.5	93.5	93.2	93.4
top 5	95.3	96.2	95.9	95.8	94.2	95.7	95.6	95.8

Table 2. Speaker adaptation phrase recognition rates (%)

initial HMnet	single Gaussian (speaker selection)				6-mixture (speaker-independent)			
# training phrases	5	10	16	24	5	10	16	24
top 1	85.5	87.7	88.4	88.7	85.1	86.9	88.0	89.5
top 3	94.8	96.3	96.4	96.5	94.7	95.1	96.0	96.7
top 5	96.5	97.6	97.7	97.7	96.8	96.9	97.3	97.9

HMnet when the adaptation data were few (less than 16 phrases). This is because only the selected speaker cluster suitable for use as the input speaker was used. If the adaptation data are sufficient, roughly more than 24 phrases, the 6-mixture HMnet may be advantageous owing to its large degree of freedom. This experiment's results suggest changing the mixture number according to the amount of adaptation training data as a guide to a dynamic speaker adaptation [11].

In addition, this framework was able to drastically reduce the likelihood map computation with its speaker selection mechanism. This mechanism combined with the speaker adaptation was able to minimize the likelihood map computation to one-sixth its previous size without any degradation to recognition performance.

6 CONCLUSION

We have developed a novel speaker parsing method for speaker-independent speech recognition. This method offers a speaker selection mechanism to efficiently adapt acoustic models to the input speaker. Since the speaker-consistency principle best demonstrates its effect with a large number of training and test speakers, a small-scale experiment may not fully exploit this principle. Nevertheless, even in our small-scale experiment the new method outperforms the conventional speaker-independent method.

The speaker parsing method can reduce the likelihood map calculation because all hypotheses of low-ranking speakers may be pruned at an early stage by the beam search and therefore those speakers' density functions do not need to be calculated. Moreover, this framework can drastically reduce the likelihood map computation by utilizing its speaker selection mechanism. After the speaker selection/adaptation process, the likelihood map computation becomes only one-mth of the original. Although the single Gaussian HMnet had one-sixth degree of freedom, it outperformed the 6-mixture HMnet when the adaptation data are few.

In addition, if noisy environments are regarded as speaker variations, this framework, here called "noise-consistent parsing", can also be applied to noisy speech recognition. It can also be applied to speaker-identification/-verification by exploiting the speaker selection mechanism.

ACKNOWLEDGMENTS: The authors would like to thank Dr. Yamazaki, President, ATR Interpreting Telecommunications Research Laboratories, and Dr. Sagisaka, head of the lab's Department 1, for their continuous support. We are also grateful to Mr. Kosaka, Mr. Takami (Victor), Mr. Miyazawa (EPSON) and all the other members of Department 1 for their discussions and encouragement.

REFERENCES

[1] M. Bates, R. Bobrow, P. Fung, et al.: "Design and Performance of HARC, the BBN Spoken Language Understanding System", *Proc. ICSLP-92*, pp.241–244 (1992).

[2] X. Huang, F. Alleva, M.-Y. Hwang and R. Rosenfeld: "An Overview of the SPHINX-II Speech Recognition System", *ARPA Workshop* (1993).

[3] K. Ohkura, M. Sugiyama and S. Sagayama: "Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs", *Proc. ICSLP-92*, pp.369–372 (1992).

[4] P. Niyogi and V. W. Zue: "Correlation Analysis of Vowels and their Application to Speech Recognition", *Proc. Eurospeech-91*, pp.1253–1256 (1991).

[5] A. Imamura: "Speaker-Adaptive HMM-Based Speech Recognition with a Stochastic Speaker Classifier", *Proc. ICASSP-91*, pp.841–844 (1991).

[6] V. Abrash, H. Franco, M. Cohen, N. Morgan and Y. Konig: "Connectionist Gender Adaptation in a Hybrid Neural Network / Hidden Markov Model Speech Recognition System", *Proc. ICSLP-92*, pp.911–914 (1992).

[7] A. Nagai, J. Takami and S. Sagayama: "The SSS-LR Continuous Speech Recognition System: Integrating SSS-derived Allophone Models and a Phoneme-context-dependent LR Parser", *Proc. ICSLP-92*, pp.1511–1514 (1992).

[8] J. Takami and S. Sagayama: "A Successive State Splitting Algorithm for Efficient Allophone Modeling", *Proc. ICASSP-92*, pp.I-573–I-576 (1992).

[9] T. Kosaka, J. Takami and S. Sagayama: "Rapid Speaker Adaptation Using Speaker-Mixture Allophone Models Applied to Speaker-Independent Speech Recognition", *Proc. ICASSP-93*, pp.II-570–II-573 (1993).

[10] T. Kosaka and S. Sagayama: "Tree-Structured Speaker Clustering for Fast Speaker Adaptation", *Proc. ICASSP-94*, pp.I-245–I-248 (1994).

[11] T. Kosaka, E. Willems, J. Takami and S. Sagayama: "A Dynamic Approach to Speaker Adaptation of Hidden Markov Networks for Speech Recognition", *Proc. Eurospeech-93*, pp.363–366 (1993).