SPEECH RECOGNITION BASED ON ACOUSTICALLY DERIVED SEGMENT UNITS

Toshiaki Fukada†, *Michiel Bacchiani*††, *Kuldip K. Paliwal*†††*and Yoshinori Sagisaka*†

†ATR Interpreting Telecommunications Research Laboratories, Japan ({fukada,sagisaka}@itl.atr.co.jp)
††Department of ECS Engineering, Boston University, USA (bacchian@raven.bu.edu)
†††School of ME, Griffith University, Australia (K.Paliwal@me.gu.edu.au)

ABSTRACT

This paper describes a new method of word model generation based on acoustically derived segment units (henceforth ASUs). An ASU-based approach has the advantages of growing out of human pre-determined phonemes and of consistently generating acoustic units by using the maximum likelihood (ML) criterion. The former advantage is effective when it is difficult to map acoustics to a phone such as with highly co-articulated spontaneous speech. In order to implement an ASU-based modeling approach in a speech recognition system, we must first solve two points: (1) How do we design an inventory of acoustically-derived segmental units and (2) How do we model the pronunciations of lexical entries in terms of the ASUs. As for the second question, we propose an ASU-based word model generation method by composing the ASU statistics, that is, their means, variances and durations. The effectiveness of the proposed method is shown through spontaneous word recognition experiments.

1. INTRODUCTION

In speech recognition, current successful approaches are mainly based on context-dependent phone modeling with distribution clustering techniques. These approaches achieve 90% recognition accuracy for unlimited-vocabulary read Wall Street Journal speech and 97% accuracy for a roughly 5000 word vocabulary spontaneous human-computer database query task. However, in the case of human-human dialog utterances, for example, the Switchboard corpus, we can only get about 50% accuracy even if state-of-the-art acoustic models are used. At ATR, we are collecting spontaneous human-human dialog utterances[1]. Our current system achieves only about a 60% word correct rate for speaker-independent models and 65% for speaker-adapted models. This suggests that a radical shift in modeling is needed to handle some of the phenomena found in spontaneous speech.

Statistical acoustic models have been studied mainly based on phone units which have been pre-determined independently of real acoustic characteristics of spoken utterances. This pre-determination of phone units causes serious mismatches between input speech characteristics and recognition unit characteristics invoked by corresponding phone sequences, especially when it is applied to highly co-articulated spontaneous speech.

To cope with these mismatches, we combined two advances proposed in previous work [2][3]. The first is the use of acoustically derived segment units (ASUs), which was an active research topic in the late 1980's. Secondly, the ASUs are represented by stochastic segment trajectory models where the trajectories can be specified with an arbitrary regression order [3].

In order to implement an ASU-based modeling approach in a speech recognition system, we must solve two problems: (1) How do we design an inventory of acoustically-derived segmental units and (2) How do we model the pronunciations of lexical entries in terms of the ASUs. For the first question, Bacchiani et al. have proposed automatic generation schemes [4]. As for the second question, if we have a large number of word speech to be recognized, we can construct an ASU-based statistical word model[5][2]. In general, however, it is difficult to construct such database especially for large vocabulary systems. To overcome this problem, we have developed an ASU-based word model generation method by composing the ASU statistics.

We start with a brief explanation on ASU generation in section 2. Section 3 then presents a method of mapping scheme between ASUs and a lexicon. Word recognition experiments on speaker-dependent spontaneous speech are described in section 4. Section 5 discusses some issues in the proposed method and shows other approaches.

2. ACOUSTICALLY DERIVED UNIT GENERATION

2.1 Polynomial Trajectory Models

Each ASU can be represented by stochastic segment trajectory models where the trajectories is specified with an arbitrary regression order. This modeling is the same as that proposed by Gish et al.[3] except that we adopt this modeling not for phones but for ASUs.

2.2 Unit Design

The ASU generation is carried out as follows. First, acoustic segmentation is done as an initialization of the unit design procedure. Second, the segments resulting from the acoustic segmentation are clustered to form an initial inventory of ASUs. As the acoustic segment boundaries obtained by the first step are sub-optimal for the initial inventory, iterative re-estimation is done. We have confirmed experimentally that only a few iterations are quite enough[6].

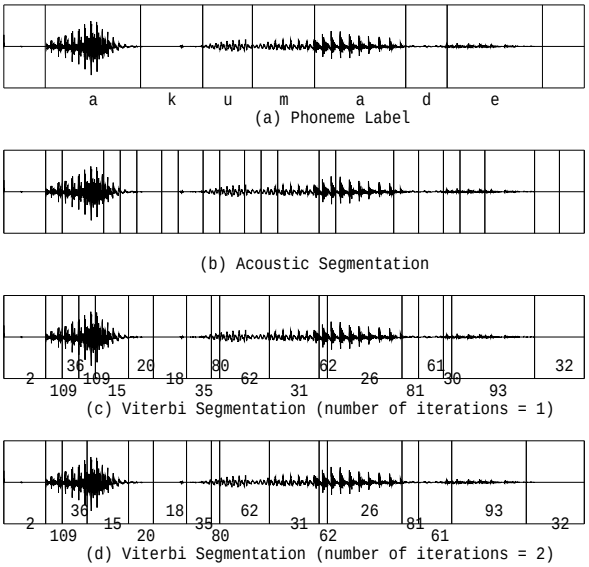

Figure 1: Example of ASU-based segmentation.

2.3 Segmentation Example

Figure 1 shows an example of ASU-based segmentation. The boundaries include (a) hand-labeled phoneme boundaries, (b) acoustically segmented initial boundaries, (c) Viterbi segmented boundaries by initial ASUs (calculated from the acoustic segmentation) and (d) Viterbi segmented boundaries using a secondly calculated ASU set via Viterbi segmentation.

Using this iterative algorithm, we have confirmed experimentally that (1) the test set likelihood using the ASUs is higher than that based on traditional phone models and (2) the likelihood as a function of the iterations on the testing data increases monotonously[4].

3. LEXICAL MAPPING

In ASU-based speech recognition, the main question to be answered is how to represent words in the recognition vocabulary in terms of an appropriate sequence of ASUs. Several techniques have been proposed for the case in which a large number of utterances for each vocabulary word are seen in the training set[2][5]. However, no method has been proposed for unseen words. To cope with this problem, we propose an ASU-based composition method which enables the production of lexicon-based word models. These word models are made in a three-step process: (1) phoneme level composition, (2) word level composition, and (3) hybrid composition using the results of (1) and (2). These steps are outlined below.

3.1 Phoneme-based ASU Composition

First, the training data is acoustically segmented by performing a Viterbi segmentation given the ASUs from the training data as described in section **2.** Then, using a hand-labeled or automatically labeled time-aligned phonemic transcription, segmented ASU sequences are divided into phonemic units. A phoneme model is generated through the following steps:

1. Choose a representative sample $\tilde{O}$ for segment alignment from M samples $O(i)(i = 1, \ldots, M)$ which has the corresponding phoneme as a central phoneme:
$$\tilde{O} = \underset{i}{\operatorname{argmax}} \sum_{m=1}^{M} P(O(m), O(i)) \tag{1}$$
where $P(\cdot)$ indicates the probability between samples that can be calculated using ASU statistics (i.e., the ASU means, the ASU variances and the durations of the time-aligned ASU transcription) and weights according to the contextual coincidence.
2. Align the segments of $O(i)$ to those of $\tilde{O}$ by dynamic programming with their means.
3. For ASUs $(k = 1, \ldots, K)$ which are assigned to one ASU of the representative sample, perform temporal composition to obtain the mean $x_{ph}(m)$ and the variance $\sigma_{ph}(m)$ of the phonemic unit:
$$x_{ph}(m) = \frac{\sum_{k=1}^{K} l_k(m) x_k(m)}{\sum_{k=1}^{K} l_k(m)} \tag{2}$$
$$\sigma_{ph}(m) = \frac{\sum_{k=1}^{K} l_k(m) \left[\sigma_k(m) + \{x_k(m) - x_{ph}(m)\}^2\right]}{\sum_{k=1}^{K} l_k(m)} \tag{3}$$
where $x_k(m)$, $\sigma_k(m)$ and $l_k(m)$ are the mean, variance and duration of the k-th ASU segment, respectively.
4. Perform contextual composition by merging ASUs for each phonemic segment using their means $x_{ph}(m)$ and variances $\sigma_{ph}(m)$:
$$\bar{x}_{ph} = \frac{\sum_{m=1}^{M} w_{ph}(m) x_{ph}(m)}{\sum_{m=1}^{M} w_{ph}(m)} \tag{4}$$
$$\bar{\sigma}_{ph} = \frac{\sum_{m=1}^{M} w_{ph}(m) \left[\sigma_{ph}(m) + \{x_{ph}(m) - \bar{x}_{ph}\}^2\right]}{\sum_{m=1}^{M} w_{ph}(m)} \tag{5}$$
where $\bar{x}_{ph}$ and $\bar{\sigma}_{ph}$ are the composed mean and variance for the phoneme of the lexicon, and $w_{ph}(m)$ is a weight determined based on the contextual coincidence.

Currently, six phonemes (three left and three right) are considered as a phonetic context. The duration statistics is

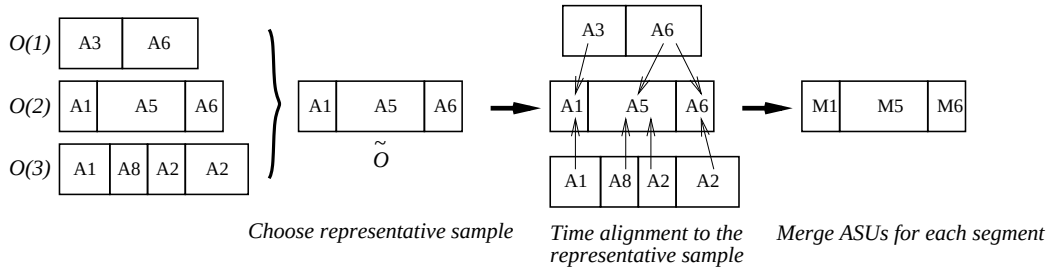

Figure 2: Example of phoneme model composition scheme. A3 represents an ASU whose code is three and M1 represents a composed segment code.

calculated from phonetic durations weighted by the contextual coincidence and represented as a Gaussian distribution. Figure 2 illustrates a process of generating a phoneme model from three samples (i.e., $O(1), O(2)$ and $O(3)$). Finally, a word model can be constructed by concatenating the phoneme models. Each word model is made by its own composed phoneme model generation and concatenation.

3.2 Word-based ASU Composition

Now, we extend the phoneme-based composition method to the word level to generate more precise word model. To this end, in step 1 described in **3.1**, we choose a representative vocabulary speech sample $\tilde{O}_{wd}$ for segment alignment from M samples $O_{wd}(i)(i = 1, \ldots, M)$ whose utterance is the same. Then a word model can be generated in the same way as steps 2 and 3. Note that phoneme context dependent weighting is not performed here.

$$\bar{x}_{wd} = \frac{1}{N} \sum_{n=1}^{N} x_{wd}(n) \tag{6}$$

$$\bar{\sigma}_{wd} = \frac{1}{N} \sum_{n=1}^{N} \left[\sigma_{wd}(n) + \{x_{wd}(n) - \bar{x}_{wd}\}^2 \right] \tag{7}$$

To obtain reliable whole word models, the number of word utterances in the training set needs to be sufficiently large. In general, however, the recording of such homogeneous amounts of speech data is both impractical and unthinkable especially for large vocabulary recognition system. Therefore, we take an approach here to generate reliable (i.e. robust and precise) whole word model by composing this word model and the phoneme concatenated word model. The whole word model can be obtained by DP alignment using their composed means. Under this composing, word sample (i.e. M) dependent weighting w_N is performed.

$$\bar{x}_{word} = \frac{\bar{x}_{ph} + w_N \bar{x}_{wd}}{1 + w_N} \tag{8}$$

$$\bar{\sigma}_{word} = \frac{\bar{\sigma}_{ph} + w_N \bar{\sigma}_{wd}}{1 + w_N} + \frac{(\bar{x}_{ph} - \bar{x}_{word})^2 + w_N (\bar{x}_{wd} - \bar{x}_{word})^2}{1 + w_N} \tag{9}$$

We expect that the additional use of vocabulary speech data will enable one to construct a robust and precise word model according to the number of lexical utterances in the training data.

Table I: Speech analysis.

Sampling freq.	16 kHz
Preemphasis	0.98
Window	Hamming window, 25.6 ms
Feature parameter	10-dimensional MFCC + energy
Frame shift	10 ms

Table II: ASU generation conditions.

Acoustic segmentation	Distortion threshold	1.0
	Regression order	0
	Distortion measure	Mahalanobis
Clustering	Codebook size	120
	Distortion measure	ML
	Covariance type	Diagonal

4. EXPERIMENTS

4.1 Conditions

To confirm the baseline performance of the ASU composition method, we performed 200-word recognition experiments on speaker-dependent spontaneous speech, using the "ATR Travel Arrangement Corpus"[7][1]. The conditions for the feature parameter extraction and ASU generation are listed in Table I and Table II, respectively. To generate ASUs and word models, 237 spontaneous speech utterances by one male speaker were used. Phoneme label information, which was needed for the word model generation step, was obtained by performing automatic segmentation using speaker dependent HMMs. The context dependent weighting w_{ph} discussed in **3.1** was selected as $w_{ph} = i + j + k$, where i and j are the number of coincidences of the left and right phoneme context, respectively. If both i and j were greater than or equal to 1, k was set to 20. Otherwise, k was set to zero. Word sample dependent weighting w_N was set to $0.1n$, where n is the number of word samples, so as to be weighted equally between whole word model which was constructed from ten samples and phoneme concatenated model.

4.2 Results

As our reference system, we used HMM-based context dependent phoneme models[8] with 400 states and a single Gaussian density function per state. The model topology generation and training were performed using the same 237 spontaneous speech utterances.

The recognition experiments showed that the recognition rate of the phoneme-based ASU composition method described in **3.1** was 80.5%, while conventional recognition rate was 80.0%. Furthermore, by word-based ASU composition method described in **3.2**, the recognition rate was improved to 82.0%. These results support our approach and superior performance for continuous spontaneous speech recognition.

5. FURTHER MODELING OF LEXICAL MAPPING

5.1 Problems of the ASU Composition Method

With the ASU composition method described in section **3.**, each ASU is transformed by composition to produce lexicon dependent word models. As a result, the model parameters increase in proportion to the size of the lexicon, yet robust and precise word models are still generated.

In the word recognition experiment, the regression order of the mean trajectories was set to zero. Though higher regression order modeling would produce better recognition performance than 0-th order modeling, the composition procedure becomes complicated.

To overcome these problems, several approaches that use the original ASU statistics as pronunciation networks, and not based on composing, could exist[6]. We present here a neural network(NN) based approach for pronunciation network generation.

5.2 Pronunciation Network Generation based on Neural Networks

A sequence of cepstral feature vectors representing an input utterance to be recognized is acoustically segmented by performing a Viterbi segmentation given the ASUs. If we consider modeling this ASU sequence O_t $(1 \leq t \leq M)$ as an observation and phoneme sequence $q_t = j$ $(1 \leq j \leq N)$ defining the hidden states, then we can define a forward variable $\alpha_t(j)$ in a manner similar to the forward procedure of the conventional HMM:

$$\begin{aligned}
\alpha_t(j) &= P(O_1, \ldots, O_t, q_t = j|\lambda) \\
&= \sum_{i=1}^{N} P(O_1, \ldots, O_t, q_{t-1} = i, q_t = j|\lambda) \\
&= \sum_{i} P(O_1, \ldots, O_{t-1}, q_{t-1} = i|\lambda) \\
&\quad \cdot P(O_t, q_t = j|O_1, \ldots, O_{t-1}, q_{t-1} = i, \lambda) \\
&= \sum_{i} \alpha_{t-1}(i) \\
&\quad \cdot P(O_t|O_1, \ldots, O_{t-1}, q_{t-1} = i, q_t = j, \lambda) \\
&\quad \cdot P(q_t = j|O_1, \ldots, O_{t-1}, q_{t-1} = i, \lambda) \qquad (10)
\end{aligned}$$

where λ is a set of models of ASUs. Assuming that O_t is independent of observations $O_1, \ldots, O_{t-2}$ and that $q_t = j$ is independent of time and $O_1, \ldots, O_{t-1}$, then Eq. (10) can be rewritten as

$$\alpha_t(j) = \sum_{i} \alpha_{t-1}(i)\ P(O_t|O_{t-1}, i, j, \lambda)\ a_{ij} \qquad (11)$$

where a_{ij} is the cost of going from the i-th phoneme to the j-th phoneme. Using Eq. (11), we can readily derive a Viterbi style decoding algorithm.

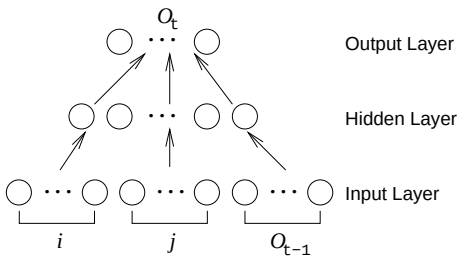

Figure 3: An example of an NN structure.

To calculate the probability $P(O_t|O_{t-1}, i, j, \lambda)$ in Eq. (11), we introduce an NN paradigm as below. First, using the training data which has phoneme label and ASU label information, we assign ASUs to phonemes. Then the NN which has the structure shown in Figure 3 can be trained by a back-propagation scheme. The recognition result of this work will be reported in the near future.

6. SUMMARY

In this paper, we have presented a new method of ASU-based word model generation. An ASU-based approach has the advantages of growing out of human pre-determined phonemes and of consistently generating acoustic units by using the ML criterion. The effectiveness of the proposed method is shown through spontaneous word recognition experiments. Furthermore, we pointed out the problem of the proposed method and presented the further modeling method based on neural networks.

REFERENCES

[1] A. Nakamura et al.: "Japanese Speech Databases for Robust Speech Recognition," *Proc. ICSLP-96*, 1996.

[2] K. Paliwal: "Lexicon-building methods for an acoustic sub-word based speech recognizer," *Proc. ICASSP-90*, pp. 729–732, 1990.

[3] H. Gish and K. Ng: "A Segmental Speech Model with Applications to Word Spotting," *Proc. ICASSP-93* pp. II-447–II-450, 1993.

[4] M. Bacchiani, M. Ostendorf, Y. Sagisaka and K. Paliwal: "Unsupervised Learning of Non-Uniform Segmental Units for Acoustic Modeling in Speech Recognition," *Proc. IEEE ASR Workshop 95* pp. 141-142, 1995.

[5] T. Svendsen, F. Soong and H. Purnhagen: "Optimizing Baseforms for HMM-based Speech Recognition," *Proc. EUROSPEECH-95* pp. 783–786, 1995.

[6] M. Bacchiani, M. Ostendorf, Y. Sagisaka and K. Paliwal: "Design of a Speech Recognition System Based on Acoustically Derived Segmental Units," *Proc. ICASSP-96*, 1996.

[7] T. Morimoto et al.: "A speech and language database for speech translation research," *Proc. of ICSLP94*, pp. 1791-1794, 1994.

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