An Iterative Approach to Model Merging for Speech Pattern Discovery

Lei Wang*, Eng Siong Chng* and Haizhou Li*†
* Nanyang Technological University, Singapore
E-mail: {WANG0161@e.ntu.edu.sg, aseschng@ntu.edu.sg} Tel: +65-6790 4964
† Institute for Infocomm Research, Singapore
E-mail: hli@i2r.a-star.edu.sg Tel: +65-6408 2773

***Abstract*—This paper introduces a novel approach to automatically discover recurrent speech patterns from multi-speaker corpus without *a priori* knowledge. The proposed approach is based on the sub-word acoustic units and it iteratively concatenates the most-likely joint sub-word units to produce a longer acoustic unit till our proposed stop criterion is satisfied. Among the resulting acoustic units, the units with the most stable number of occurrences are selected as the lexicon. The proposed approach has been applied to automatically discover English words from TIDIGIT corpus. The experimental results measured by F1 score showed the proposed approach can effectively detect and extract the recurrent patterns. This technique can be used for lexicon generation from an unknown speech corpus or in audio content summarization.**

I. Introduction

Automatic speech pattern discovery remains a difficult problem with continued interest [1], [2]. This research topic will benefit the study of human's cognitive system on language acquisition, for example, how does an infant learn its mother language? How can researchers create lexicon of unknown languages?

According to the psychological study [3], the infant English learners, at the age of 10.5months, show their sensitivity to statistical regularities and other cues when they identify word boundaries from fluent speakers. Inspired by such findings, computer scientists started effort to explore techniques to discover speech patterns in an unsupervised manner.

Previous research [4], [5] has studied the discovery of fundamental acoustic units. The detailed literature survey is discussed in Session II-A. Among the techniques, Acoustic Segment Model (ASM) [4] was proposed to automatically model the sub-word units from unknown speech corpus. Although ASM can analyze basic acoustic units of an unknown language, it is unable to assign meanings to those units to create a lexicon. Researchers were therefore motivated to examine speech patterns on the semantic level. Another approach is based on the conventional dynamic time warping (DTW). As the DTW is applied directly on the front-end output, it may suffer from noise and multi-speaker effect in speech signal. To improve robustness, modeling approach can be examined to archive the same goal.

Inspired by Jusczyk's study [3], a novel approach built on information theory and ASM is proposed to demonstrate infants' language acquisition mechanism in this paper. Our approach can be applied to explore unknown languages by extracting the frequently occurred words or phrases hence it offers a solution to build the lexicon of an unknown language.

The remainder of the paper is organized as follows: Section II briefly reviews related literature for speech pattern discovery and it consists of two parts: acoustic units modeling and speech pattern detection, Section III proposes our approach to detect recurrent patterns using sub-word units, Section IV reports the experimental results and finally, we conclude in Section V.

II. Related Work

During the past decade, speech recognition techniques model the acoustic units based on given transcripts and the recognition process can be interpreted as a path searching mechanism using both the acoustic and language models to find the best hypothesis of the sentence read given the input signal. However, such techniques cannot be applied to the corpus of which the manual transcript is unavailable or the language is unknown. This motivates researchers to explore novel approaches to discover speech pattern automatically.

A. Acoustic Units Modeling

From late 1980s to early 1990s, researchers realized conventional HMM approach needs to overcome challenges in dealing with spontaneous speech due to acoustic variability. One of their arguments was to question whether the phonemes should be considered as the smallest units in speech. Many researchers put effort to explore the substituted units such as sub-words, syllables and other data-driven units from speech data. Such studies can be considered as pioneering work of speech pattern discovery, and Ostendorf summarized these previous work as "beads-on-a-string" model [6]. The following paragraphs review some of the studies in details.

The researchers in [5] examined method to build Markov models for isolated words by concatenating predefined fenone models. A fenone is known as a sub-phone unit in speech and it represents a single frame in their work. In their approach, a small amount of transcribed speech data is used to train 200 prototypes of fenones. To build word models, each frame of input frame is assigned a pre-defined fenonic label based on nearest neighbor criteria. The fenonic sequence is then used as a reference to construct the word model by concatenating the

fenone Markov models. The recognition is carried out using dynamic programming (DP). However, such method can only be applied to single speaker corpus due to the limitation on predefined prototypes.

Unlike modeling fenones in isolated words speech, Takami and Sagayama [7] proposed Successive State Splitting (SSS) to study an optimal representation for a set of acoustic units in continuous speech automatically. By applying SSS, a hidden Markov network (HM-Net), which is a single HMM with a number of trained states with optimal parameters in an optimal topology, can be generated. In each iteration of SSS, the state with the largest divergence in HM-Net will be split into two states in either temporal or contextual domain, depending on the reduction in likelihood of all samples. Each state sequence of HM-Net can be interpreted as a phone or sub-word unit. As an extension of SSS, Singer and Ostendorf [8] considered maximum likelihood criterion for selection of the state and the way to be split so that the HM-Net can grow towards to global optimal. Recently, Varadarajan *et al.* [9] improve the efficiency of SSS by simultaneously splitting all the existing states into 4 followed by merging back states with less occupancy. They also further proposed to use finite state machine as a transducer to automatically assign meaningful label to each state sequence hence SSS could be applied to continuous speech recognition in single speaker corpus.

In another attempt, a data-driven approach, namely Acoustic Segment Model (ASM), was proposed to characterize fundamental speech sounds for speech recognition in [4]. The ASM models have the similar topology as conventional HMMs of speech recognizer, however, its training process does not require transcript of speech data. Instead, ASM approach iteratively decodes unlabeled training data and re-estimate the ASM models using the latest decoding output by maximum likelihood criterion. Hence, the iterative training process can update ASM models towards the best representation of the data and each resulting ASM model represents a sub-word unit. These sub-word units have been further explored to create lexicon in [10]. The idea of ASM was also subsequently extended by Li *et al.* [11] to train universal phoneme models for the language identification task.

B. Speech Pattern Detection

Extending the acoustic unit modeling approach, recent research of speech pattern discovery focuses on detecting meaningful patterns such as words instead of sub-word units. For example, a project named "ACORNS" (Acquisition of Communication and Recognition Skills) was launched in Europe in 1996. The objective of the project is to develop a computational model to acquire human verbal communication behavior. Their primary goal for the research work is to automatically build a word inventory for 10 words.

Different approaches have been explored to discover speech patterns in unsupervised manner. One effective approach is to first segment speech into tokens and then detect the repeating patterns from token sequences. For instance, [12] explored the local alignment algorithm to discover similar portions from multiple phoneme sequences in topic spotting application. Another approach examined in [13] applied non-negative matrix factorization (NMF) on the soft counts of the bigrams to explore the recurrent speech patterns. However, NMF cannot provide the order information of the obtained acoustic units.

Dynamic Time Warping (DTW) is another approach which has been promoted in speech pattern discovery. Recently, Park and Glass [1] proposed the segmental DTW, a variant of conventional DTW, to detect repeating speech patterns between spoken utterances pairs. The method divides the entire search space into sub-space for patterns detection using dynamic programming (DP). All the detected patterns are then clustered into groups based on distance so that frequently occurred patterns can be captured. In an application of [1], researchers in [2] extended segmental DTW to obtain extractive summary by evaluating the sentences that contain those keywords. Other work such as [14] also detected repeating keywords using variant of DTW in call-center application. The DTW-based methods are applied directly on the front-end output hence it may suffer from noise and multi-speaker effect.

III. Iterative Approach to Model Merging

A. The Proposed Framework

Our proposed framework is inspired by [5], and we propose to use an iterative approach to model merging that grows sub-word models towards word/phrase models. Figure 1 shows the overall framework of our proposed approach. The system begins with a set of Acoustic Segment Models (ASMs) [4] to represent sub-word acoustic units. Based on the decoding sequence using these initial ASMs, information theory is applied to evaluate each pair of the acoustic units so that the pair which is the most likely to be together can be identified. Subsequently, a new and longer acoustic unit is formed by merging the selected acoustic pair. The process continues iteratively till a specific stop criterion is satisfied.

After the training process, a pattern selection criterion is used to rank the resulting patterns hence the top-ranked patterns can be selected to create a lexicon. The following sections describe each module in details.

B. Acoustic Segment Model (ASM)

In this section, the ASM models are briefly described. The ASM [4] was introduced to learn acoustic units in an unsupervised manner. Li *et al.* [11] further introduced a bootstrapped ASM training procedure for universal phonetic tokenization. This unsupervised approach does not require transcription for training. Instead, vector quantization is first applied to produce pseudo labels to initialize the training procedure. The procedure to generate ASM models is discussed in the following paragraph.

The features extracted from the speech data are the MFCC features as well as their first and second order time derivative. The ASM training process is applied to automatically generate the a set of M HMMs in the following manner:

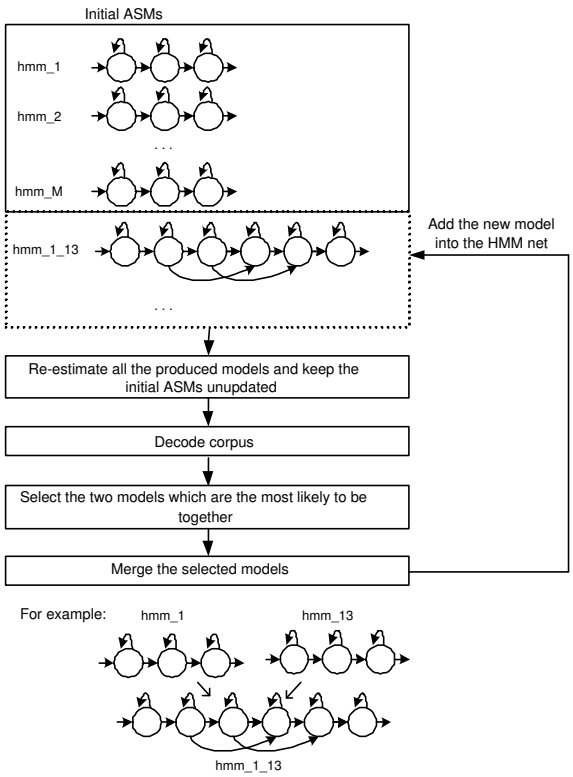

Fig. 1. A diagram of the iterative approach to model merging.

Step 1) Divide the speech utterances into short segments using the speech segmentation method described in [15].

Step 2) Cluster the segments into M clusters using k-means and assign cluster identities to the segments.

Step 3) Create M HMM models. Adapt these M HMM models using the training corpus with the labeled cluster identity found by Step 2.

Step 4) The trained M HMMs are used to decode the training corpus.

Step 5) The HMMs are re-estimated using the new labels found after Step 4.

Step 6) Repeat 4 -5 until convergence.

C. Models Merging Criteria

At the end of each iteration of the training process, the corpus is decoded by the new HMM models using a free phone loop, which will be described in the subsequent sessions. The selection of the models to be merged is based on bigram counts of the decoding output.

Following the information theory, different criteria have been successfully applied to textual word segmentation. We begin by studying the transitional probability, which is shown in Eq. 1, to decide word boundaries [16], and mutual information is also used in word boundary detection [17], [18] as shown in Eq. 2.

$$p(b|a) = \frac{p(a,b)}{p(a)} \tag{1}$$

$$I(a;b) = \log_2 \frac{p(a,b)}{p(a)p(b)} \tag{2}$$

where a and b represent two characters/words/sub-words in a symbolic sequence.

We found that the above two mentioned criteria are not robust for speech pattern discovery. One possible reason could be the size of speech corpus is much smaller than the pure textual data. To include additional context information, we modify the above transitional probability to include the reverse bigram counting as shown in Figure 2. The representation of our proposed criterion is

$$S(a;b) = \log_2 \frac{p(a,b)^2}{p(a)p(b)} \tag{3}$$

$$= \log_2 p(b|a)p(a|b) \tag{4}$$

$$\widehat{a;b} = \arg\max S(a;b) \tag{5}$$

where, a and b are two acoustic units in the decoding sequence.

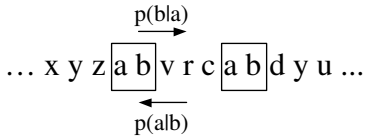

Fig. 2. Symmetric bigram counting.

The new model will be formed by concatenating the two existing HMM models selected using Eq. 5, and it will be added to the set of models as illustrated in Figure 1. Hence, all the existing models and newly formed model are retained in this process.

This model merging process will stop either when $S(\widehat{a;b}) < \gamma$ where γ is a user defined threshold or the bigram term $\widehat{a;b}$ has been chosen before. A high γ value will prevent a semantically meaningful speech pattern (e.g. words and short phrases) from being formed; While a low setting will reduce the recall of some meaningful patterns as it might divide the same meaningful pattern into sub-classes.

D. Patterns Selection Criteria

The previous merging process will generate a set of HMMs which represent different acoustic patterns such as sub-word units, syllables, words and short phrases. Words and short phrases are linguistically meaningful patterns and the ability to identify them automatically is of great interest. We proposed the following strategy to find meaningful patterns.

The meaningful patterns should satisfy two conditions: (i) high frequency in the corpus, and (ii) stable models, i.e. the pattern model should have low chance to be further merged with other patterns. Condition (i) can be simply evaluated by counting the number of occurrences of a pattern in the decoding output. While condition (ii) is not easy to be measured as we do not have prior knowledge about the corpus. However, we may achieve this through investigating the stability of the patterns. Once a meaningful pattern such as a complete word is produced, it will not be easily merged with other

patterns according to our model merging criteria as unformed patterns will rank higher in the merging criteria. Hence, second order statistics of the number of occurrences in each training iteration can be used to evaluate condition (ii) in our case.

IV. Experiments

To verify our proposed iterative approach to model merging, we examine how it discovers English digital words from TIDIGIT corpus. The TIDIGIT corpus has a small vocabulary size that consists of only 11 English words: digits 0–9 and "oh". There are 8440 utterances that spoken by male and female speakers and the data is clean. Each utterance contains either isolated word or connected words, and the average length of the utterances is about 7sec. We designed different experiments by partitioning the corpus into gender groups and developement/testing sets. Both the development and testing sets of each gender have 2110 utterances.

The front-end uses the 13-D MFCC feature with its delta and double double that extracted by HTK with window size 25msec and hop size 10msec. We applied our proposed approach to male and female utterances separately. In both of the experiments, a set of $M = 16$ HMMs were generated using the ASM training process.

To evaluate the performance of the proposed approach, the detected speech segments are aligned to the time-labeled references. A 20-frame error interval is allowed at each side of the boundaries of a reference word. A majority voting strategy is then used to assign label to each detected pattern model so that recall and precision can be calculated for each pattern.

TABLE I
Average F1 score on development and testing sets.

	Dev Set	Test Set
Male	0.83	0.80
Female	0.84	0.71

According to patterns selection criteria, the top 10 ranked patterns are selected in each experiment. The best overall performance that our proposed approach can achieve on development sets is shown in Table I. From the development sets, the following parameters are derived: 1) the user defined threshold γ is set to -2.737 for both male and female development sets; 2) The word insertion penalty, which is used to add additional value to each model when it transits from end of one word to the start of the next [19], is set to -46 and -38 for male and female data, separately. The parameter settings are then applied on the testing sets and the average F1 scores are also shown in Table I.

Table II shows a few examples of mapping between the HMM sequences to English digit words. To further analyze the HMM models, the reference phoneme boundaries are also used in this study. For instance, in the female data set experiment, pattern "hmm_1 hmm_13 hmm_16" can be mapped to word "one", and pattern "hmm_3 hmm_8 hmm_14 hmm_16" can be mapped to word "seven". We realize that model "hmm_16" occurs at the end of both patterns. As the words "one" and "seven" have the same last phoneme, i.e. /n/, we hypothesize the phonetic equivalent of the original ASM model "hmm_16" is /n/. Our hypothesis can be verified by aligning the HMM sequences with the reference phoneme boundaries. Similarly, the ASM model "hmm_3" can be found to represent the fricative phonemes which include /s/ and /f/. During the models merging process, ASM model "hmm_3" is merged by different context and further adapted to the particular phonetic equivalent.

TABLE II
Examples of mapping model sequences to words.

HMM model sequence	References
hmm_10 hmm_12 hmm_14	two
hmm_1 hmm_13 hmm_16	one
hmm_3 hmm_1 hmm_7	four
hmm_3 hmm_8 hmm_14 hmm_16	seven
hmm_10 hmm_7 hmm_2	three
hmm_3 hmm_11 hmm_5 hmm_3	six

V. Conclusion

In this paper, a novel approach is proposed to automatically learn recurrent patterns such as words or phrases from label-free multi-speaker speech corpus. The proposed approach is based on acoustic modeling so that it is more robust in multi-speaker environment. Information theory is used to guide the discovering process of new speech patterns. The proposed approach has been successfully applied to study digital words from TIDIGIT database. This work can be further applied to identify keywords from untranscribed speech or generate lexicon for unknown languages.

References

[1] A. S. Park and J. R. Glass, "Unsupervised pattern discovery in speech," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 186–197, Jan. 2008.

[2] X. Zhu, G. Penn, and F. Rudzicz, "Summarizing multiple spoken documents: finding evidence from untranscribed audio," in *Proc. the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Singapore, Aug. 2009, pp. 549–557.

[3] P. W. Jusczyk and R. N. Aslin, "Infants' detection of the sound patterns of words in fluent speech," *Cognitive Psychology*, vol. 29, no. 1, pp. 1–23, 1995.

[4] C.-H. Lee, F. K. Soong, and B.-H. Juang, "A segment model based approach to speech recognition," in *Proc. ICASSP '88*, New York, NY, 1988, pp. 501–504.

[5] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer, and M. A. Picheny, "A method for the construction of acoustic markov models for words," *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 4, pp. 443–452, Oct. 1993.

[6] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proc. IEEE ASRU Workshop*, Keystone, Colorado, 1999, pp. 79–84.

[7] J. Takami and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," in *Proc. ICASSP '92*, San Francisco, CA, Mar. 1992, pp. 573–576.

[8] H. Singer and M. Ostendorf, "Maximum likelihood successive state splitting," in *Proc. ICASSP '96*, Atlanta, GA, May 1996, pp. 601–604.

[9] B. Varadarajan, S. Khudanpur, and E. Dupoux, "Unsupervised learning of acoustic sub-word units," in *Proceedings of ACL-08: HLT*, Columbus, Ohio, USA, 2008, pp. 165–168.

[10] K. K. Paliwal, "Lexicon-building methods for an acoustic sub-word based speech recognizer," in *Proc. ICASSP '90*, Alburquerque, NM, Apr. 1990, pp. 729–732.

[11] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 271–284, Jan. 2007.
[12] P. Nowell and R. K. Moore, "The application of dynamic programming techniques to non-word based topic spotting," in *Proc. Eurospeech '95*, Madrid, Spain, Sept. 1995, pp. 1355–1358.
[13] V. Stouten, K. Demuynck, and H. Van hamme, "Discovering phone patterns in spoken utterances by non-negative matrix factorization," *IEEE Signal Processing letters*, vol. 15, pp. 131–134, 2008.
[14] M. Cevik, F. Weng, and C.-H. Lee, "Detection of repetitions in spontaneous speech in dialogue sessions," in *Proc. Interspeech 2008*, Brisbane, Australia, Sept. 2008, pp. 471–474.
[15] L. ten Bosch and B. Cranen, "A computational model for unsupervised word discovery," in *Proc. Interspeech 2007*, Antwerp, Belgium, Aug. 2007, pp. 1481–1484.
[16] J. R. Saffran, R. N. Aslin, and E. L. Newport, "Statistical learning by 8-month-old infants," *Science*, vol. 274, no. 5294, pp. 1926–1932, 1996.
[17] K. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, Mar. 1990.
[18] K. Church, W. Gale, P. Hanks, and D. Hindle, "Using statistics in lexical analysis," in *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pp. 115–164. Erlbaum, 1991.
[19] S. Young et al., *The HTK Book (for HTK Version 3.2.1)*, Cambridege University Engineering Department, 2002.