Singapore DSO National Labs Project – FINAL REPORT

Keyword Recognition and Correction Based on Utterance Verification and Knowledge Integration of Acoustic-Phonetic Features

(Submitted by Chin-Hui Lee, Georgia Institute of Technology, chl@ece.gatech.edu)

1. Introduction

Recognition of keywords embedded in extraneous and unconstrained speech is of practical importance in many real-world applications, ranging from recognizing five call types in operator services [Wilpon90] to understanding a few thousand key phrases in ill-formed spoken queries in spontaneous dialogue interactions [Kawahara98]. The abilities to model competing non-keywords and reject extraneous speech events are two critical factors in determining the performance of keyword recognition systems.

If the number of keywords of interest is small, such as in the case of voice recognition call processing (VRCP) operator service [Wilpon90], we can collect whole-word spoken examples to train individual models, one for each keyword, and a single filler, also called "garbage collection", unit to model all non-keyword speech. On the other hand when the number of keywords is over a few dozens the distinction between keywords and competing non-keyword event becomes less obvious. This often results in more keyword misrecognition and non-keyword false-alarm errors. Such a problem becomes even more serious when sub-word units, such as phones instead of whole-words, are used to represent both key-phrases and non-keyword speech. One way to alleviate the above difficulties is to use large vocabulary continuous speech recognition (LVCSR) systems to recognize both keyword and non-keywords. However the performance of such LVCSR-based keyword recognition algorithms depends heavily on the quality of the acoustic and language models used in the LVCSR systems. Since we often do not have enough text data to train reliable domain-specific language models an alternative is to train language models for key-phrases and non-keyword fillers separately [Kawahara98], and combine them to represent the entire language models needed in the LVCSR-based keyword recognition systems. When the available set of training data is really limited, like in most keyword recognition scenarios, it is usually not feasible to train good non-keyword acoustic and language models in order to obtain satisfactory performances.

In this project we adopt a two-pass keyword recognition and correction approach based on utterance verification [Sukkar96, Lee97, Rahim97] and integration of acoustic-phonetic features for rescoring of candidates in a list or a lattice [Sukkar97, Lee04, Li05, Lee07, Ma07, Siniscalchi09]. The first pass is to obtain keyword hypotheses with an *n*-best candidate list or lattice followed by a second pass that corrects potentially wrong candidates through integration of acoustic-phonetic knowledge sources for rescoring. This second-pass keyword rescoring component can utilize detailed acoustic models or knowledge sources that are required to distinguish subtle differences, such as nasal endings in "teens" in "fourteen" versus "forty" or fricative beginning and ending in "five" versus "nine". Acoustic phonetic knowledge sources [Lee04] are especially useful because they can be modeled well with a relatively small amount of speech data as shown in the recently proposed automatic speech attribute transcription (ASAT) approach to automatic speech recognition [Lee07]. High performance detection of speech attributes, such as manner of articulation has been recently reported [Li05-2]. When the knowledge scores produced by these acoustic-phonetic detectors are incorporated into HMM-based

likelihood scores, significant accuracy improvement has been obtained in place where articulatory inconsistency cannot be corrected by detailed phone models [Siniscalchi09].

2. Proposed Statement of Work in 2010-2012

We propose a 24-month project to start in April 1, 2010. We assume the current DSO keyword detection module and its extensions will be used in the first pass of the keyword recognition system. Three subprojects for the second pass of the system are planned, one for each DSO personnel involved in the proposed project, namely: (1) implement minimum classification error (MCE) [Juang97] and minimum verification error (MVE) [Lee97] training and prepare generalized likelihood ratio (GLLR) plots [Tsao05] for all the sub-word units of interest; (2) designing utterance verification (UV) subsystem [Sukkar96] and high-performance statistical tests [Lee97] to reject non-keywords falsely detected in the first pass mentioned above; and (3) experimenting with integration of speech attribute detectors [Li05-2, Siniscalchi09] with acoustic-phonetic features [Lee07] to enhance rejection of false alarms and correct misrecognized keywords [Sukkar97, Li05, Ma07, Siniscalchi09]. We will use the N-TIMIT corpus for all modeling and recognition experiments. The same system development methodology is expected to be extended to any telephone-based keyword recognition task. We will therefore adopt short-term exchange visits by getting the people performing the above subprojects to visit Georgia Tech as visiting scholars to experiment with different tasks and different databases. The budget to support these visiting stays will be allocated separately. The PI will also visit Singapore twice per year to monitor project progresses. Teleconference meetings will be scheduled on a regular basis to discuss project milestones as well.

3. Schedule and Milestones

We expect the proposed two-pass framework to lead to performance improvement over the current one-pass approach. The three subtasks of the proposed project are listed as follows, with their milestones shown in Table 1 with a total effort of three-person years.

1. **Discriminative training of phone models and separation characterization among models:** studying issues relate to modeling of phone separations and enhancing them through minimum classification/verification error (MCE/MVE) discriminative training.
2. **Utterance verification:** designing word-level and phone-level verification tests to reduce false alarms. The results from subtask 1 will be highly relevant for this subtask.
3. **Integration of acoustic-phonetic features:** designing attribute detectors to reduce false alarms and correct some detected words not following acoustic-phonetic constrains. This subtask can follow closely with the results from the ASAT paradigm. Detectors based on artificial neural networks and hidden Markov models will be experimented.
4. **Two-pass system integration:** modifying the current first-pass to generate multiple lists or lattice of detected hypotheses and applying techniques developed in subtasks 2 and 3 to the second pass of the system. This requires constant testing and evaluation.

Table 1 Sub-tasks and project milestones over the two years of the SoW

SubTask	Year 1				Year 2			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
1	*	*			*	*		
2		*	*	*	*	*	*	
3			*	*	*	*	*	*
4	*	*	*	*	*	*	*	*

4. Summary

The system developed under the proposed project had led to new and effective techniques for correcting confusable words not easily distinguishable by conventional phone models in medium-vocabulary keyword recognition. They had been trialed in the field and achieved better performance and enhanced robustness. Deployment of the system is under way. In the meantime better technologies had been developed along the way when the team was engaged in the OpenKWS2013 and OpenKWS2014 evaluations sponsored by National Institute of Standards and Technologies (NIST). These new techniques will also be folded into the system to be deployed later.

5. Reference

[1] J. G. Wilpon, L. R. Rabiner, C.-H. Lee and E. R. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models," *IEEE Trans. Acoustic, Speech and Signal Proc*., Vol. ASSP-38, pp. 1870-1878, Nov. 1990.

[2] R. A. Sukkar and C.-H. Lee, "Vocabulary Independent Discriminative Utterance Verification for Non-Keyword Rejection in Subword Based Speech Recognition," *IEEE Trans. on Speech and Audio Proc*., Vol. 4, No. 6, pp. 420-429, Nov. 1996.

[3] R. Sukkar, A. R. Setlur, C.-H. Lee and J. Jacob, "Verifying and Correcting String Hypotheses Using Discriminative Utterance Verification," *Speech Communication*, Vol. 22, pp. 333-342, 1997.

[4] M. Rahim and C.-H. Lee, "String-Based Minimum Verification Error (SB-MVE) Training for Speech Recognition," *Computer, Speech and Language*, Vol. 11, No. 2, pp. 147-160, April 1997.

[5] B.-H. Juang, W. Chou and C.-H. Lee, "Discriminative Methods for Speech Recognition," *IEEE Trans. on Speech and Audio Proc*., Vol. 5, No. 3, pp. 257-265, May 1997.

[6] C.-H. Lee, "A Unified Statistical Hypothesis Testing Approach to Speaker Verification and Verbal Information Verification," invited paper in *Proc. COST Workshop on Speech Technology in the Public Telephone Network: Where are we today?*, pp.62-73, Greece, September 1997.

[7] T. Kawahara, C.-H. Lee and B.-H. Juang, "Key-Phrase Detection and Verification for Flexible Speech Understanding," *IEEE Trans. on Speech and Audio Proc*., Vol. 6, No. 6, pp. 558-568, Nov. 1998.

[8] C.-H. Lee, "From Knowledge-Ignorant to Knowledge-Rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition," Plenary Talk, *Proc. ICSLP*, Jeju, South Korea, October 2004.

[9] J. Li, Y. Tsao and C.-H. Lee, "A Study on Knowledge Source Integration for Candidate Rescoring in Automatic Speech Recognition," *Proc. ICASSP-2005*, Philadelphia, March 2005.

[10] Y. Tsao, J. Li and C.-H. Lee, "A Study on Separation between Acoustic Models and Its Applications," *Proc. InterSpeech-2005*, Lisbon, Portugal, September 2005.

[11] J. Li and C.-H. Lee, "On Designing and Evaluating Speech Event Detectors," *Proc. Interspeech*, Lisbon, Portugal, September 2005.

C.-H. Lee, "An Overview on Automatic Speech Attribute Transcription (ASAT)," *Proc. Interspeech*, Antwerp, Belgium, August 2007.

[12] C. Ma and C.-H. Lee, "A Study on Word Detector Design and Knowledge-Based Pruning and Rescoring," *Proc. Interspeech*, Antwerp, Belgium, August 2007.

[13] S. M. Siniscalchi, T. Svendsen, and C.-H. Lee, "A Phonetic Feature Based Lattice Rescoring Approach to LVCSR," *Proc. ICASSP*, Taipei, Taiwan, April 2009.