A Study on Word Detector Design and Knowledge-based Pruning and Rescoring

Chengyuan Ma and Chin-Hui Lee

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332, USA
{cyma, chl}@ece.gatech.edu

10.21437/Interspeech.2007-427

Abstract

Detection of speech attributes, phones and words is a key component of a detection-based automatic speech recognition framework in the automatic speech attribute transcription project. This paper presents a two-stage approach, keyword-filler network method followed by knowledge-based pruning and rescoring, for detection of any given word in continuous speech. Different from conventional keyword spotting systems, both content words and function words are considered in this study. To reduce the high miss, a modified grammar network for word detection is proposed. Then knowledge sources from landmark detection, attributes detection and other spectral cues were combined together to remove the unlikely putative segments from the hypothesized word candidates. This study has been evaluated on the WSJ0 corpus under matched and mismatched acoustic conditions. When comparing with the conventional keyword spotting system, we found the proposed word detector greatly improves the detection performance. The figure-of-merits for content and function words were improved from 48.8% to 61.5%, and 22.3% to 33.1% respectively.

Index Terms: word detection, knowledge-based

1. Introduction

Research on automatic speech recognition (ASR) has achieved great success in the last several decades. Nevertheless, some challenging problems still exist. Among them, detection of out of vocabulary (OOV) words and out of grammar events are the major limitations. Part of the reason causing these difficulties is that our knowledge about phonetics, phonology and linguistics has not been fully integrated into the ASR system [1] [2]. It has been shown that an integration of additional knowledge sources is beneficial to improving the robustness of ASR systems [3]. Recently, an automatic speech attribute transcription (ASAT) paradigm has been proposed to fully incorporate knowledge sources into the ASR system through a bottom-up detection of fundamental speech units followed by a knowledge integration process [1] [2]. Under the ASAT framework, detection of speech attributes, phones and words is a key component of a detection-based ASR system. At the lower level, the articulatory manner and place attributes, landmarks and some salient spectral cues, which are robust to environment and speaker variations, have been investigated with statistical modeling and signal processing techniques [4] [5] [6]. At a higher level, the phone detectors based on both acoustic and knowledge features have been studied extensively [7]. Meanwhile, some efforts have been done on combining the lower-level attributes into the higher-level acoustic events like phones [8]. At the word level, conventional keyword spotting (KWS) systems directly work on acoustic features and only focus on content words, which are generally long words [9] [10] [11]. In a recent study, knowledge-guided detector for a single digit and a detection-based ASR system for digits has been demonstrated [12]. A more detailed survey and summary on the research efforts in this area can be found in [13].

In this study, we focus on the detector network design for any single word with both data-driven method and knowledge-based pruning and rescoring. The detection-based ASR system was designed to work in both domain-dependent and domain-independent scenarios. So some cross-corpus evaluations were conducted to simulate domain-independent testing. The other part of this study is knowledge-based pruning and rescoring. Our study shows that using the proposed grammar network and knowledge-based pruning and rescoring strategies, the detection performance for both content and function words can be greatly improved using both domain-dependent and domain-independent phone models.

2. Single word detector design

A single word detector is essentially a KWS system that aims at detecting any single word in continuous speech. In conventional KWS research, function words are generally treated as stop-words that were excluded from the keyword list. In a detection-based ASR system, any single word, including both content and function words needs to be considered. KWS has been studied extensively in spoken document indexing and retrieval, spoken message understanding, and speech surveillance applications [9] [10] [11]. In general, KWS methods can be categorized into two groups. The first one is based on large-vocabulary continuous speech recognition (LVCSR), either in vocabulary-dependent word lattice or vocabulary-independent phone lattice. The other one is the keyword-filler network based method. In LVCSR based methods, language model plays an important role. Previous research showed that the word lattice based method has the best performance [11]. Nevertheless, the disadvantage of the word lattice based method is that the vocabulary and the language are fixed in advance. On the other hand, the phone lattice based method is more flexible and vocabulary-free. However, its performance is usually worse than the keyword-filler network based method due to the low phone recognition accuracy [11]. Hence, in this study, we choose the keyword-filler network based method as our baseline system. It is also vocabulary-independent and has better performance than phone lattice based method.

The confidence metric is of crucial importance in a KWS

system. Because keyword likelihood scores often change with the keyword (depending mostly on word length), some normalization is necessary for confidence computation. The most commonly used confidence measures are the likelihood ratio, local posterior probability and their variants [9]. Given an observation sequence of the hypothesized word segment, $O = \{o_t, t = t_s, \cdots, t_e\}$, the frame-normalized log-likelihood ratio (LLR) is defined as follows:

$$LLR(O) = \frac{1}{t_e - t_s} \sum_{t=t_s}^{t_e} \log \frac{p(o_t|\Lambda_k)}{p(o_t|\Lambda_b)} \quad (1)$$

Here, t_s and t_e are the beginning and ending time indices of a detected segment, and Λ_k and Λ_b are keyword model and background model, respectively.

2.1. Filler and background model selection

Filler models are used to fill the non-keyword speech intervals, while the background models are used to calculate the confidence measures for putative keywords. The most widely used filler model is the phone-loop model. In this way, an appropriate bonus should be put in the keyword path. Otherwise, the unconstrained phone-loop will absorb the whole speech utterance. In our preliminary study on filler model selection, 5 broad phonetic class models (vowel, nasal, stop, etc.) are used as filler model and provide better performance than phone-loop model. This is because with the less precise filler models, the keyword get more chances to occur even under an inappropriate penalty weight. The background model used for score normalization is trained using the acoustic observations from all phones.

2.2. Grammar network design

Generally, a straightforward grammar network (as shown in Figure 1) is used for keyword spotting [11]. Filler models are placed in parallel to the keyword model, which is composed from the sub-word unit models. In matched acoustic condition and with an appropriately chosen penalty, this network can generate very good result with high precision and recall. However, the keyword spotting performance depends on penalty selection greatly. In fact, once the models have been chosen, the penalty selection is the only factor that change the operating point in a KWS system.

Figure 1: *Conventional grammar network for KWS.*

We proposed a grammar network that is different from the state-of-the-art KWS systems [11] for arbitrary word detection (see Figure 2). This network has been demonstrated successful in our previous study on detection-based digits recognition experiments [12]. Undoubtedly, less misses will occur with this network. Moreover, it doesn't rely on the empirical penalty parameters.

3. Knowledge-based pruning and rescoring

In the two-stage detection approach, the first step involves generating a list of putative hits which specify the boundary of word segment, and a confidence score associated with it. Inevitably, with both the conventional and the proposed network, there will be many false alarms in the hypothesized word segments. It agrees with our expectation, i.e., less misses and more false alarms. It leaves a large space for other knowledge sources to be used. In the following, knowledge sources, pruning and rescoring strategies will be presented.

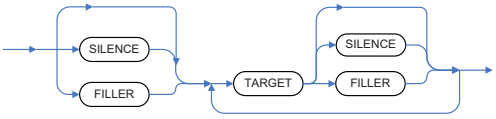

Figure 2: *Proposed grammar network for word detector.*

3.1. Knowledge sources

Many acoustic, phonetic and linguistic knowledge sources have been studied intensively in previous decades [5] [7]. In this papers, the landmark detection and distinctive features are employed in detailed analysis [4] [5]. In addition, an artificial neural network (ANN) based phone recognizer [14] was used to provide phone information and manner attributes for each small segment within a putative word segment. This ANN based phone recognizer uses long-span cepstral features and has achieved the best phone recognition performance on TIMIT corpus and several other applications [14]. The detailed knowledge used in this study are shown as below:

durational information: Both the inherent and contextual durational information provide important perceptual cues. For instance, voiceless fricatives (e.g., /s/, /f/, /sh/, /th/) are 40ms longer than their voiced counterpart, (/z/, /v/, /zh/, /dh/) [15].

landmarks: Six landmarks have been detected for each segment. For example, [+g] and [-g] indicate the turning on/off the glottal vibration respectively. [+s] and [-s] mean a closure/release of nasal or /l/. [+b] and [-b] indicate the burst of stop sounds [4] [5].

manner attributes: The manner attributes have been shown to be robust to environment and speaker variations.

formant transition pattern: Formant transition pattern for some vowels are easily recognized and reliable.

phone confusion matrix: In addition to manner attributes, phone confusion probability provides more detailed discrimination within each broad phonetic class. For example, for vowel /ae/, it is more likely to be misrecognized as /eh/ than others vowels.

degree of voicing: It will be used to distinguish between voiced and voiceless sounds.

3.2. Pruning

From some observations, some false alarms are easy to detect. For example, a simple minimal durational constraint works well in practice. Another simple method is to use the Levenshtein distance (edit distance) between the detected and expected manner attribute sequences. Our experiments showed that by choosing loose thresholds with regard to the number of sounds in a detected word, nearly 50% of the false alarms can be removed without reducing the detection rate. In addition, this result is consistent with our expectation that longer words are easier to

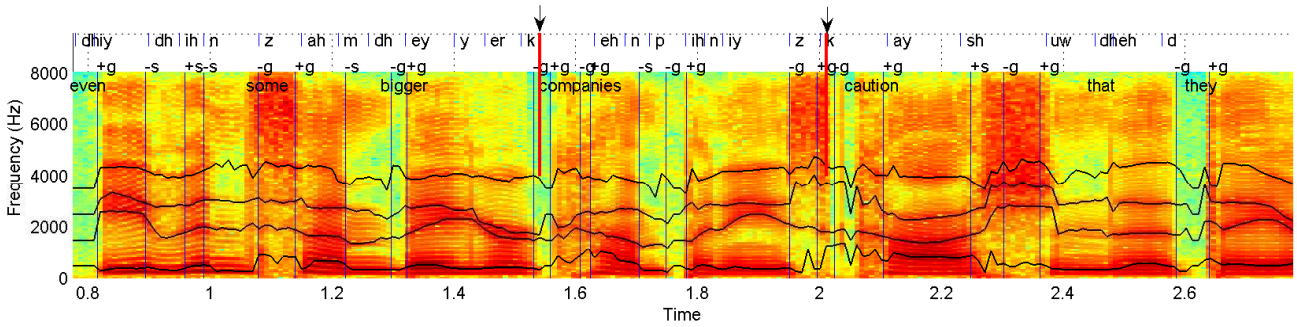

Figure 3: *Knowledge sources for detected word.*

be correctly detected. Model based pruning can also be implemented. Furthermore, signal processing based pruning strategies are more desirable because they are robust across different acoustic environments. Detailed analysis is designed for the hard confusions, and it is based on the above-mentioned knowledge sources.

Figure 3 shows an utterance from the WSJ0 testing set. The topmost in the figure is the recognized phone sequence using the ANN based phone recognizer trained from the TIMIT corpus. The second sequence is the landmarks described in Section 3.1. The speech segment between the two arrows is a hypothesized segment for word "company". The vertical lines indicate the location of each landmark. For example, a [-s] landmark shows the start of a nasal sound. The region between a [+g] and [-g] landmark is a voiced sound. For this voiced region, we can further analyze the formant transition pattern. Similar procedure can be applied to voiceless region indicated by [-g] and [+g] landmarks.

3.3. Rescoring

Pruning is the way of making hard decision. By pruning, some hypothesized word segments are removed completely. However, in some cases, there is no strong evidence or salient feature to support pruning. Rescoring, as a soft decision strategy, is a good choice in knowledge combining. Since the confidence measure has a great impact on the performance evaluation. After pruning, the remaining hypothesized word candidates are rescored by a simple rescoring strategy described in [6]. It's a weighted linear combination of LLR and the phone posterior probability generated by the ANN phone recognizer. By rescoring, the detected words are re-ranked according to their new confidence scores in figure-of-merit (FOM) computation.

4. Experiment setup and result analysis

All the evaluation experiments were carried out on the WSJ0 corpus. Both the WSJ0 training set (7132 sentences from 84 speakers) and the TIMIT training set (3696 sentences from 462 speakers) were used in acoustic modeling of context independent monophone models, broad phonetic class model and background model for cross-corpus evaluation. The WSJ0 testing set (Nov92 non-verbalized 5k closed set) consists 330 sentences from 8 speakers. A conventional procedure is used for front-end processing. To conduct cross-corpus evaluation and reduce the channel effects, every element of the feature vector has been normalized with zero-mean and unit-variance [12].

The keywords, including both content and function words, are randomly selected from the original 5k WSJ vocabulary. 30 content and 20 function words have been chosen and they are listed in Table 1. The cut-off frequency was set to 8 when selecting keywords to ensure a reliable evaluation.

Table 1: *List of content words and function words.*

content words	analyst average bank bill company day dollar exchange first good hundred issue market million month one order percent plan point real rose said seven share state stock thousand time year
function words	after also but by from had has have he in may of that the their then this which will would

4.1. Performance measure

The performance of a KWS system is usually measured using receiver operating characteristic (ROC) curves and figure-of-merit (FOM) [16]. FOM is an upper-bound estimate of word spotting accuracy averaged over 1 to 10 false alarms per hour. It's the area under the part of the ROC curve with false alarms from 1 to 10 per hour. In practice, we have little interest in the area beyond 10 false alarms per hour. A keyword is considered successfully detected if the mid-point of the hypothesis fell within the reference time interval. All the hypothesized keywords are sorted with respect to their confidence score, and the probability of detection at each false alarm rate was then computed. An average FOM over all keywords is used as the overall performance measure. Generally speaking, there will be more false alarms with more keywords in the keyword list.

4.2. Comparative experiments

Several comparative experiments have been conducted. The first one was to evaluate the performance of the conventional KWS system under matched (WSJ0 monophone models) and mismatched (TIMIT monophone models) acoustic conditions. It's clear that there is a big performance gap between content words and function words. Even in matched acoustic condition, FOM of content words (48.8%) is two times larger than that of function words (22.3%). The performance drop caused by the acoustic mismatch agrees with our expectation. FOM decreased from 48.8% and 22.3% in matched condition to 42.6% and 18.4% in mismatched condition.

The second experiment (as shown in Table 3) was to conduct knowledge-based pruning and rescoring on the output of a conventional KWS system. We can see for both content words

Table 2: *FOM for conventional method.*

	WSJ0 Model	TIMIT Model
Function Words	22.3%	18.4%
Content Words	48.8%	42.6%

and function words, under both matched and mismatched conditions, the performance has been improved a lot. FOM increased from 48.8% to 58.9% for content words in matched condition and similar results are achieved for function words. It manifests the effectiveness of the knowledge-based pruning and rescoring strategy.

Table 3: *FOM for conventional method with pruning.*

	WSJ0 Model	TIMIT Model
Function Words	29.5%	25.1%
Content Words	58.9%	54.7%

The third experiment (as shown in Table 4) was to conduct knowledge-based pruning and rescoring on the output of the proposed network and filler model selection. For content words, FOM increased from 58.9% in Table 3 to 61.5% in Table 4. This small improvement should attribute to the new network structure. Comparing with the result in Table 2, the performance improvement is significant. For content words, FOM increased from 48.8% to 61.5%.

Table 4: *FOM for proposed method with pruning.*

	WSJ0 Model	TIMIT Model
Function Words	33.1%	29.7%
Content Words	61.5%	58.3%

The experiment results showed in Table 4 is comparable with other state-of-the-art KWS systems [10] [11]. In [10], FOM is 73.8% on a 20 keyword task with triphone model and in [11], FOM is 64.5% on a 17 keyword task with monophone models. It's clear that both the proposed grammar network and the knowledge-based pruning and rescoring strategy are very effective, even with less detailed acoustic model (monophone models) and under mismatched condition (TIMIT models). However, the performance difference between content words and function words are still very large. The most frequently occurred function words are short words and often are part of other words, like "of" in "offer". Language model and other linguistic knowledge will be helpful for dealing with these short function words.

5. Summary and future work

In this paper, we proposed a two-stage approach for arbitrary word detection in continuous speech. Following a modified keyword-filler network detection process, knowledge sources like landmarks, manner attributes, durational information have been explicitly incorporated into the system. The performance of single word detection has been greatly improved when comparing with the conventional KWS system. The pruning and rescoring strategies described in this paper is somewhat straightforward and mostly rule based. They are far away from being perfect. In future studies, this single word detector will be embedded into a detection-based large-vocabulary continuous speech recognition system.

6. Acknowledgement

This work was partially supported by the NSF ITR grant, IIS-04-27413.

7. References

[1] Lee, C.-H., "On Automatic Speech Recognition at the Dawn of the 21st Century," *IEICE Trans. Inf. & Syst.*, pp. 377–396, 2003.

[2] Lee, C.-H., "From Knowledge-ignorant to Knowledge-rich Modeling: A New Speech Research Paradigm for Next Generation Automatic Speech Recognition," in *Proc. ICSLP*, 2004.

[3] Kirchhoff, K., "Combining Articulatory and Acoustic Information for Speech Recognition in Noisy and Reverberant Environments," in *Proc. ICSLP*, 1998.

[4] Liu, S. A., "Landmark Detection for Distinctive Feature-based Speech Recognition," *JASA*, pp. 3417–3430, 1996.

[5] Stevens, K. N., "Toward a Model for Lexical Access Based on Acoustic Landmarks and Distinctive Features," *JASA*, pp. 1872–1891, 2002.

[6] Siniscalchi, S., Li, J. and Lee, C.-H., "A Study on Lattice Rescoring with Knowledge Scores for Automatic Speech Recognition," in *Proc. InterSpeech*, 2006.

[7] Ali, A., Spiegel, J., Mueller, P., Haentjens, G. and Berman, J., "An Acoustic-phonetic Feature-based System for Automatic Phoneme Recognition in Continuous Speech," in *Proc. ISCAS*, 1999.

[8] Morris, J. and Fosler-Lussier, E., "Combining Phonetic Attributes using Conditional Random Fields," in *Proc. InterSpeech*, 2006.

[9] Rose, R. C. and Paul, D. B., "A Hidden Markov Model Based Keyword Recognition System," in *Proc. ICASSP*, 1990.

[10] James, D. A. and Young, S. J., "A Fast Lattice-based Approach to Vocabulary Independent Wordspotting," in *Proc. ICASSP*, 1994.

[11] Szöke, I., Schwarz, P., Matějka, P., Burget, L., Karafiát, M., Fapšo, M. and Černocký, J., "Comparison of Keyword Spotting Approaches for Informal Continuous Speech," in *Proc. InterSpeech*, 2005.

[12] Ma, C., Tsao, Y. and Lee, C.-H., "A Study on Detection Based Automatic Speech Recognition," in *Proc. InterSpeech*, 2006.

[13] Matthews, B., et al., "Detection-Based ASR in the Automatic Speech Attribute Transcription Project," submitted to *InterSpeech*, 2007.

[14] Schwarz, P., Matějka, P. and Černocký, J., "Towards Lower Error Rates in Phoneme Recognition," in *Proc. TSD*, 2004.

[15] Chung, G., Hierarchical Duration Modelling for a Speech Recognition System, S.M. thesis, MIT Department of Electrical Engineering and Computer Science, 1997.

[16] NIST, "The Road Rally Word-spotting Corpora (RDRALLY1)," NIST Speech disc 6-1.1, 1991.