Keyword Detection for Spontaneous Speech

Weifeng Li[†], Aude Billard[†], and Hervé Bourlard[†,‡]
[†]*Swiss Federal Institute of Technology, Lausanne (EPFL), Switzerland*
[‡]*Idiap Research Institute, CH-1920, Martigny, Switzerland*
Email: {weifeng.li}@epfl.ch

Abstract

This paper presents a system for keyword detection in spontaneous speech. Keywords are predefined through a set of acoustic examples provided by the users. Keyword detection proceeds in two steps: keyword searching and verification. To address the problem of using the same phoneme models in both keyword and filter models, we propose to remove the phoneme models included in the keyword model from the filter models. In order to reduce the false alarms caused by keyword searching step, dynamic time warping (DTW) based template matching and Gaussian Mixture Models (GMM) are proposed. Our keyword detection experiments demonstrate the effectiveness of the proposed methods by yielding improved detection performance compared to the baseline system.

1. Introduction

Information retrieval from spoken audio has attracted a lot of attention in the last two decades. Keyword spotting, a special branch in continuous speech recognition, has appeared on data from telephone speech [1], air travel information [2], broadcast news task [3]. One problem of state-of-the-art keyword spotting systems is that the relevant keywords are out of vocabulary in many applications, for example when searching names, places, acronyms, etc. On the other hand, in some applications the keywords are completely predefined through a set of acoustic examples provided by the users. The task is to detect the predefined keywords and to find the exact location in the test speech.

This paper presents our keyword detection studies on such a task. We believe that an effective keyword detection system must be able, in first time to spot a keyword embedded in speech, and in second time to reject the candidate speech regions that do not include any valid keyword. The block diagram of our keyword detection system is shown in Fig. 1. In the pre-processing step, the feature parameters are extracted. Next, keyword searching is performed via Viterbi beam search by using a set of phoneme models as filter models. Finally the resulting keyword candidates (or hypotheses) are verified by a keyword model which is built from a set of acoustic examples provided by the users.

One method of keyword searching consists in introducing Finite State Grammar (FSG) according to prior knowledge to detect keyword in a whole sentence. Excellent performance can be acquired through this method by including certain speech structures in the given Finite State Grammar. The limit of FSG is its inability to cover all possible speech structures, and hence results in poor robustness on practical systems [4]. Another method, which has been proved efficient in keyword spotting and is used in this paper, is allowing filler (or garbage) models to absorb non-keywords. There are three typical approaches for absorbing non-keywords: (1) Combine all the extraneous speech regions to train one Hidden Markov Model (HMM) as a filter model; 2) Large Vocabulary Continuous Speech Recognition (LVCSR) based approach in which the garbage model only allows valid words from the lexicon; and (3) Assemble the phoneme models to establish the filler model; The first approach does not work well because the established filter model can not cover all the variabilities in test speech. Due to the use of additional linguistic constraints, the second approach was shown to improve the spotting performance [5]. Such an approach is expensive in that it requires collecting a large amount of labeled data for training LVCSR systems as well as the high computational cost [6]. Because of its ability to automatically adapt to outbursts in test speech and comparatively low computation cost, the third approach is employed in this paper.

However, the phonetic-based keyword spotting system with filter models has another drawback, due to using the same phoneme models in both keyword and filter models, which can result in degrading the recall performance of a keyword. This issue is typically addressed by using a more refined garbage model [6] or an on-line garbage model [7]. In this paper we propose to remove the phoneme models which are included in the keyword model from the filter model in the decoding network. Our experiments demonstrate that such a simple method improve the keyword recall rates significantly.

As a consequence of removing the phoneme models which are included in the keyword model from the filter model, we may obtain an increased false alarms. In order to reduce the false alarms caused by keyword searching step, two verification methods are proposed. One is Dynamic Time Warping (DTW) based template matching that gives superior

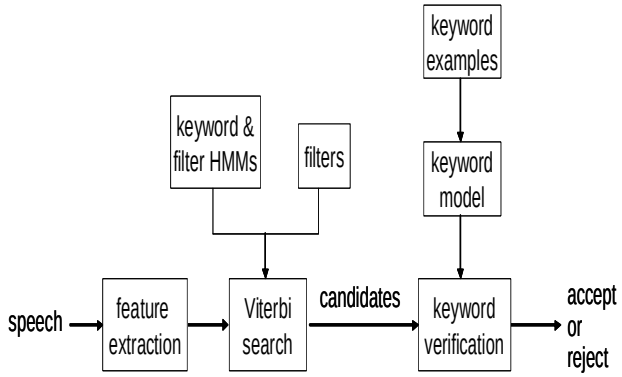

Figure 1. Configuration of our keyword detection system.

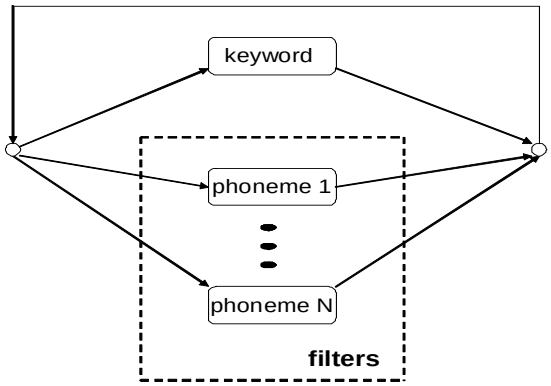

Figure 2. The keyword decoding network using filter models. The keyword model is a left-right HMM, resulting from the concatenation of the phonemes corresponding to the keyword phoneme sequence. The filters consist of phoneme models.

performance on isolated word recognition system [8]. The other is based on Gaussian Mixture Models (GMM), which has been successfully used to deal with speaker verification [9] and speaker recognition [10]. Our studies showe that these methods reduce the false alarm rates while keeping the recall rates almost intact.

This paper is structured as follows: Section 2 briefly describes experimental setup. Section 3 and 4 presents the keyword searching and verification experiments, and Section 5 concludes this paper.

2. Experimental Setup

The recognizer used in this work is a speaker independent HMM system. The modeled unit is the phoneme, each phoneme is represented by 3-state, strictly left-to-right, continuous density HMM. A word is represented by the concatenation of phoneme models. The number of probability density function (pdf) per state is determined during the training phase. We used 15 hours of speech from Conversational Telephone Speech [11] for training phoneme-level Hidden Markov Models (HMMs). All speech data are digitized into 16 bits at a sampling frequency of 8 kHz with accuracy 16 bits, pre-emphasized with $(1 - 0.97z^{-1})$. Each frame is multiplied by a hamming window with 25 milliseconds and is computed at every 10 milliseconds. 24-channel log melfilter bank analysis is then applied, which is transformed into 12 components of Mel Frequency Cepstral Coefficient (MFCC) using Discrete Cosine Transformation (DCT). Finally, we performed Cepstral Mean Subtraction (CMS) for channel compensation. Thus, 12 CMS normalized MFCCs and log-energy with corresponding delta and acceleration coefficients are used as feature vectors. In this paper, we used HTK [12] for the audio processing, feature extraction, acoustic modelling, and decoding.

We recorded five spontaneous speech sessions (2 female and 3 male speakers) in office environments. There are five keywords: 'blue', 'red', 'yellow', 'green', and 'ball'. The total number of the occurences of the keywords and non-keywords is 189 and 1,038, respectively. In each session, the five keywords are uttered 10 times each as acoustic samples (examples). The ground truth of time region of each keyword is manually labeled.

3. Keyword Searching

As described in Section 1, filter models based keyword spotting approach as shown in Fig. 2 is employed in our system, which allows for multiple occurences of each keyword in test speech. The keyword model is a left-right HMM, resulting from the concatenation of the phonemes corresponding to the keyword phoneme sequence. The filters (non-keywords) consist of phoneme models (including silence model). The searching process is based on the outputs of phoneme recognizer via Veterbi beam search [13], i.e., the optimal state sequence from the continuous audio stream as

$$S^* = \underset{S}{\operatorname{argmax}}\, P(S|\mathcal{M}, \mathcal{G}, \mathcal{O}) \quad (1)$$

where S is the candidate state sequence and $\mathcal{O}$ is the observation vector sequence. $\mathcal{M}$ and $\mathcal{G}$ represent the acoustic models (HMMs) and decoding network, respectively. For each frame, the corresponding state and log probability are obtained. A phoneme can be recognized by merging adjacent frames belonging to the same phoneme model, and a keyword can be distinguished by the concatenation of corresponding phoneme models. Thus, besides log probability scores the start and ending positions of each keyword are also recorded.

One of the problem leading to the degraded keyword spotting performance is due to the use of the same phoneme models in both the keyword and filer models. Although this issue have been addressed by other researchers (e.g.

[6]), we propose to remove the phoneme models which are included in the keyword model from the filter models for the simplicity and effectiveness. The system performance also depends on training phoneme HMMs accurately. Two types of HMMs (gender-independent and gender dependent) are trained by using 15 hours of speech from Conversational Telephone Speech [11].

Figure 3 shows the keyword spotting performance (ROC curve) using different type of HMMs and filter models. 'GI-HMM' and 'GD-HMM' denote the gender-independent and gender-dependent HMMs, respectively. 'filters1' and 'filters2' denote without and with removing the phoneme models included in the keyword model from the filter models, respectively. Correct acceptance or false acceptance corresponds to whether the middle time index of the searched hypothesis is within the ground truth of time region of the keyword or not. As expected, gender-dependent HMM performs slightly better that gender-independent HMM irrespective of the filers. By removing the phoneme models included in the keyword model from the filter models the performance is significantly improved. Performance of each keyword however varies greatly, as shown in Fig. 4 (bold lines).

The errors result from several characteristics of the test spontaneous speech data. First, there exists a mismatch between the training (telephone speech) and test data (speech recorded by a distant microphone). Secondly, background noises degrade the speech quality, thus resulting in recognition errors. Finally, various speakers (three of them are non-native speakers) and speaking styles (stress, speaking rate, articulatory habits, etc.) from the training data may cause recognition errors.

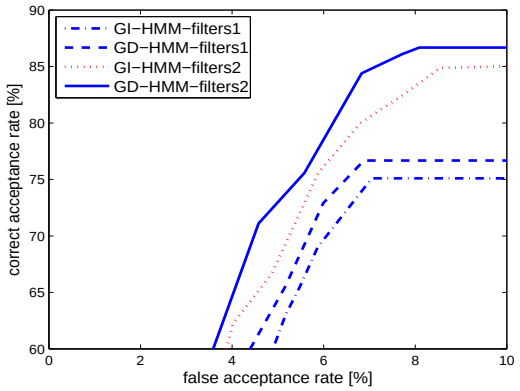

Figure 3. Receiver Operating Characteristics (ROC) curve by using different type of HMMs and filter models. 'GI-HMM' and 'GD-HMM' denote the gender-independent and gender-dependent HMMs, respectively. 'filters1' and 'filters2' denote without and with removing the phoneme models included in the keyword model from the filter models, respectively.

4. Keyword Verification

The keyword searching described above provides a set of keyword hypotheses (segmented utterances), each of which can be represented by a sequence of feature vectors (or frames) $X = \{x_1, x_2, \cdots, x_T\}$ where x_l indicates a CMS normalized MFCC feature vector at frame l. As shown in Fig. 3, 'GD-HMM-filters2' achieves the highest recall rates (86.6%) with a cost of increased false alarms compared to 'GD-HMM-filters1'. Keyword verification is an essential post-processing in our keyword detection system for it aims at rejecting the incorrectly detected keyword hypotheses while accepting as many genuine keywords as possible.

4.1. DTW based Verification

Dynamic Time Warping (DTW), which finds the optimal path from start to end for an input utterance, gives superior performance on isolated word recognition systems [8]. This simple and effective pattern matching algorithm requires less data (reference patterns) and lower computation. In our DTW based verification experiments, for each keyword we measure a distance between the detected region represent by $X = \{x_1, x_2, \cdots, x_T\}$ and the ith reference template $R^{(i)} = \{r_1^{(i)}, r_2^{(i)}, \cdots, r_P^{(i)}\}$:

$$D_i(X, R^{(i)}) = d(x_T, r_P^{(i)}) + \min D(x_T, r_P^{(i)}) \qquad (2)$$

where $d(x_T, r_P^{(i)})$ is the local distance between two vectors x_T and $r_P^{(i)}$ and $D(x_T, r_P^{(i)})$ denotes the global distance accumulated until x_T and $r_P^{(i)}$. Then an acceptance or rejection decision is made by comparing the DTW distance averaged over the reference template set with a threshold , i.e., if

$$\frac{1}{N}\sum_{i=1}^{N} D_i(X, R^{(i)}) < \delta \qquad (3)$$

satisfies, we accept this keyword candidate. Here N and δ denote the number of the reference templates and threshold, respectively.

4.2. GMM based Verification

Gaussian Mixture Model (GMM) is a parametric probability density estimation technique which has been successfully used to deal with the speaker verification [9] and speaker recognition [10]. In our experiments, each keyword is represented by a GMM

$$p(x_t|\Theta) = \sum_{k=1}^{K} w_k \mathcal{N}(x_t; \mu_k, \Sigma_k) \qquad (4)$$

where $\mathcal{N}(x_t; \mu_k, \Sigma_k)$ is the kth unimodal Gaussian density with mean vectors μ_k and covariance matrices Σ_k, and w_k

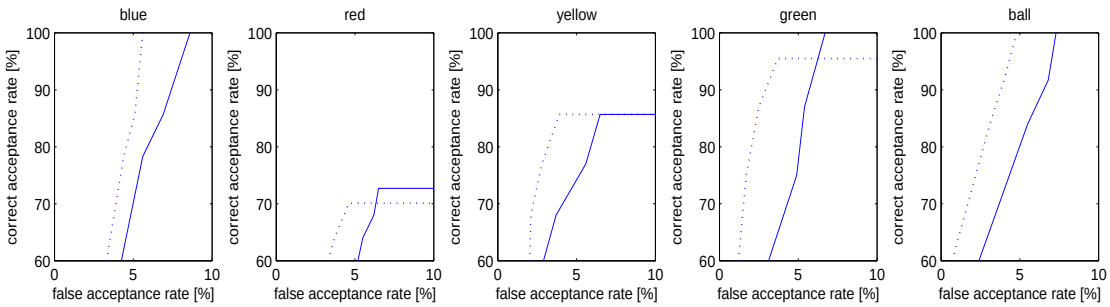

Figure 4. ROC curve for each keyword. The bold and dotted lines denote before ('GD-HMM-filters2' in Fig. 3) and after DTW based verification, respectively.

is the corresponding mixture weight. The GMM parameter set

$$\Theta = \{w_k, \mu_k, \Sigma_k\}_{k=1}^{K} \tag{5}$$

is estimated over the all the feature vectors of the keyword examples. Here K denotes the number of the Gaussian mixtures. A speech region of a keyword hypothesis represented by $X = \{x_1, x_2, \cdots, x_T\}$ is accepted as a keyword when the average log likelihood

$$\begin{aligned} L(X|\Theta) &= \frac{1}{T}\log p(X|\Theta) \\ &= \frac{1}{T}\log \prod_{t=1}^{T} p(x_t|\Theta) \\ &= \frac{1}{T}\sum_{t=1}^{T} \log p(x_t|\Theta) \end{aligned} \tag{6}$$

is larger than a threshold. Here we assume that the sequence of vectors, X, are independent and identically distributed random variables.

4.3. Experimental Results

We performed the subsequent verification experiments based on 'GD-HMM-filters2'. The number of Gaussian mixtures in GMM is set to 4 and for each keyword the thresholds are optimized experimentally. Figure 5 shows the keyword spotting performance. It is found that both verification algorithms are effective in reducing the false alarms while keeping the recall rates almost intact. Moreover DTW performs better (can reject about 50% of incorrectly detected keywords) than GMM in terms of reducing the false alarms, which can be explained by the fact that only 10 examples of each keyword are available. DTW based template matching requires less data while statistical GMM needs more training examples. Compared to the baseline system ('GI-HMM-filter1'), our final keyword detection system provides about 10% absolute improvement of true hits with much decreased false alarms. Figure 4 shows the detection performance with DTW based verifications for each keyword (dotted lines). We can see that performance of each keyword varies greatly with 'GD-HMM-filters2', but after DTW based verification the false acceptance rates for all the keywords are consistently reduced significantly. Note that some genuine speech of 'red' and 'Green' keywords could not be recovered after DTW based verifications.

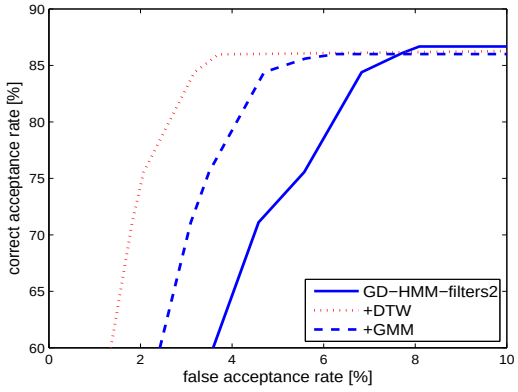

Figure 5. ROC curve after keyword verifications. 'GD-HMM-filters2' is taken from Fig. 3 for comparison.

5. Conclusions

We presented our keyword detection studies for spontaneous speech using keywords predefined by a set of acoustic examples. We proposed to remove the phoneme models included in the keyword model from the filter models to deal with the problem of using the same phoneme models in the keyword and filter models. We also proposed dynamic time warping and Gaussian mixture models to reduce the false alarms caused by the keyword searching step. Our keyword detection experiments demonstrated the effectiveness of the proposed methods in results of improved detection performance compared to baseline system. Future work lies in incorporating speaker adaptation and advanced speech processing techniques to improve the recall rates.

Acknowledgments

This work was supported by the Thought in Action (TACT) project, part of the European Union NEST-Adventure Program, and by the Swiss Science Foundation within the National Center for Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM2). The authors would like to thank Dr. John Dines and Dr. Guillermo Aradilla for the helpful discussions.

References

[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 Transactions on Acoustics, speech, and Signal Processing*, vol. 38, no. 11, pp. 1870–1878, 1990.

[2] M. Weintraub, "Lvcsr log-likelihood ratio scoring for keyword spotting," in *Proc. the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1995, pp. 129–132.

[3] S. Renals, D. Abberley, D. Kirby, and T. Robinson, "Indexing and retrieval of broadcast news," *Speech Communication*, vol. 32, pp. 5–20, 2000.

[4] C. Yining, L. Jing, Z. Lin, and L. Jia, "Keyword spotting based on mixed grammar model," in *Proc. 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing*, 2001, pp. 425–428.

[5] P. Cardillo, M. Clements, and M. Miller, "Phonetic searching vs lvcsr: How to find what you really want in audio archives," *International Journal of Speech Technology*, vol. 5, pp. 9–22, 2002.

[6] A. Manos and V. Zue, "A segment-based wordspotter using phonetic filler models," in *Proc. the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997, pp. 899–902.

[7] H. Bourlard, B. D'hoore, and J. Boite, "Aoptimizing recognition and rejection performance in wordspotting systems," in *Proc. the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1994, pp. 373–376.

[8] C. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Transactions on Acoustics, speech, and Signal Processing*, vol. 28, no. 6, pp. 623–635, 1980.

[9] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[10] S. Nakagawa, W. Zhang, and M. Takahashi, "Text-independent speaker recognition by combining speaker-specific gmm with speaker adapted syllable-based hmm," in *Proc. the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004, pp. 81–84.

[11] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Proc. the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1992, pp. 517–520.

[12] HTK, *The Hidden Markov Model Toolkit (HTK)*, version 3.4 ed. http://htk.eng.cam.ac.uk/, 2006.

[13] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. New Jersey: Prentice Hall, 1993.