Using Speech Synthesis in Keyword Spotting

Vitali Kiselov, Andre Talanov, Ivan Tampel, Marina Tatarnikova,
Speech Technology Center
St. Petersburg, Russia
{kiselev-v, andre}@speechpro.com

Abstract
The technology for unlimited vocabulary automatic keyword spotting in spontaneous Russian speech is presented in this paper. We propose a novel speech database search system. It is based on the ideas of word pattern recognition and speech synthesis. Keywords to be searched are input in the text form and corresponding speech signals are synthesized by a text-to-speech (TTS) system. These signals are used as training material for a recognizer based on the dynamic programming approach. Evaluation of the system was performed for telephone and microphone channels. In the latter case, for a limited number of keywords searched simultaneously we gain 83% of hits without speaker adaptation (false alarm at 9.3%), and 99% of hits with speaker adaptation (17.5% of false alarms). Hit rate in noisy telephone channel is 78.5%, while false alarm rate is 60%.

1. Introduction
Automatic keyword spotting (KWS) is one of the most effective means of automatic search for sound fragments in huge acoustic databases or online in sound streams. KWS finds its use in national security systems, phone services, quality control systems, broadcasting news search engines, speech message classification systems, speech filters and in many others.

Modern approaches to keyword spotting can be divided into 3 major groups [1]:

- **Word Recognition based KWS.** This method is based on isolated word recognition. Keywords are represented as indivisible units. At the training phase all acoustic representations of a keyword are combined to form its template. Matching is performed by comparing input speech segments and keyword templates. This is a one-pass method that can be used for online KWS. Within this framework foreign words can be easily included in the KWS system merely by recording enough instances for template training.

- **Phoneme Recognition based KWS.** Phoneme recognition is performed first. Keyword spotting is then performed in the phoneme lattice using phoneme posterior probabilities. The advantage of this approach is that a list of keywords must not be composed in advance. The main drawback is that the phoneme recognition error rate tends to be high when compared with the word error rate. This results in the increase in both search errors and false alarms.

- **Large Vocabulary Continuous Speech Recognition (LVCSR) based KWS.** This method is based on the use of a LVCSR system. This is a two-pass approach. Speech recognition is performed at the first pass. An LVCSR system builds a textual representation of the input speech signal. KWS is then performed not in the speech signal but in the text that is the output of a speech recognizer. This method is the most complicated due to high technological and scientific costs. The major advantage of the LVCSR based approach is that a huge sound database may be input first to a LVCSR system and the resulting text is then used to search for keywords. The textual search is several orders of magnitude faster than the search in speech.

There are a number of approaches to keyword search. Dynamic programming (DP), first proposed by J. Bridle [2], and Hidden Markov Models [3, 4] are probably the most widely known. The approach we use is the one based on dynamic time warping word recognition.

When performing word spotting, a KWS system sometimes has to deal with continuous speech of two or more speakers. Conversation topics may be arbitrary and the vocabulary used is likely to be virtually unlimited. Moreover, the language of a conversation may change. In this case, when using pattern recognition search technique, it is necessary to have arbitrary recognition templates.

The main feature of the system presented in this paper is the possibility to input keywords from the keyboard. Traditional pattern recognition systems ask a user to input via microphone and prepare a whole corpus of speech data which is then used as a corpus of templates. The main drawback of such an approach is that it is often impossible to get enough speech data for adequate template representation. To overcome this shortcoming we propose using a TTS system. This is the first time such a technique is proposed for Russian. In general, there are few papers devoted to the use of speech synthesis at the phase of KWS system. The best known pattern-based KWS is reported in [5].

2. System Overview

2.1. General Information
The system we propose consists of several fundamental blocks shown in Figure 1. Those blocks are TTS, training and matching.

*Figure 1: TTS-based KWS system*
A TTS system for Russian has recently been developed at Speech Technology Center (http://www.speechpro.com) [6, 7, and 8]. It is a state-of-the-art concatenative triphone speech synthesis system. The set of synthesis voices allows generating enough synthesized signals to form speech data templates.

The set of synthesized signals is fed to the training module where keyword templates are formed. The input audio stream is segmented into frames and analyzed by means of a filter bank. The comparison of keyword templates with stream module where keyword templates are formed. The input audio synthesized signals to form speech data templates.

The DP algorithm is used for this purpose. The DP algorithm filters the stream is segmented into frames and analyzed by means of a filter bank. The comparison of keyword templates with stream module where keyword templates are formed. The input audio synthesized signals to form speech data templates.

For keyword description and the transformation of speech signal, an algorithm of feature extraction based on IIR filtering of the second order is used:

\[ y_n = x_n + C_i y_{n-1} - r^2 y_{n-2} \quad (1) \]

where \( C_i \) is a coefficient of \( i \)th filter;
\( R \) is a pole radius (one for the whole filter bank).

The central filter frequencies are calculated as

\[ f_1 = \frac{F \times \arccos \left( \frac{C_i}{2r} \right)}{2\pi} \quad (2) \]

where \( F \) is sampling frequency.

The feature set is obtained in the following way. Input speech signal is digitized either at 11025 Hz or 8 kHz sampling frequencies. The position of the analysis frame is determined by the spectral envelope in the range of time intervals characteristic for the pitch. The pitch corresponds to pitch frequencies in the range between 80 Hz and 300 Hz. The signal singled out in a certain frame is differentiated and the filters described above are applied to it.

The output of each filter is multiplied by an amplification coefficient. This is done to ensure correspondence of the envelope to the peculiarities of human auditory system. As a next step, spectral smoothing in time domain is performed. Low-frequency filter of the second order is used for this task. A feature vector for each frame is obtained by combining neighbouring spectral components in pairs. This way a preliminary description of keyword templates and input speech signal is implemented. The dimension of feature vectors is 8 (energy and 7 spectral components).

As an additional feature, the module of after-training is available. It is used to supplement keyword templates with a speech signal (such a signal can be, for example, a speech signal prepared in advance and input with a microphone or via a telephone channel). This way the system saves keyword templates in a special database. Those templates can be rapidly extracted from the database to perform keyword search procedures.

On average, the time spent for training of one keyword (including TTS time costs) is 5-7 seconds.

2.4. Matching of Input Speech Signal Fragments and Keyword Templates

Matching of input speech stream fragments and keyword templates is done by means of the DP approach using weighted Euclidean metrics.

Input speech signal is analyzed and described as a sequence of feature vectors. It is then evaluated within a frame that corresponds to the maximum keyword template duration. The shift of the frame is chosen to be equal to the average length of a Russian phone in spontaneous speech. The frame is analyzed to check if it contains speech signal. This is done by evaluating whether spectral exceed ambient noise threshold and comparing the spectral envelope maximum with another empirical threshold.
The distance between a speech segment and keyword templates is calculated next. Within the DP approach, the path corresponding to the minimum accumulated distance when traversing between initial and final positions in both the template and the unknown signal is evaluated.

The rest of this section is devoted to the thorough description of the matching process. Let us compare two images that are characterized by two sets of vectors:

\[ X = \{ x_0, x_1, \ldots, x_i, \ldots, x_M \} \]  
\[ Y = \{ y_0, y_1, \ldots, y_i, \ldots, y_M \} \]

The difference between the vectors of two images is determined by a succession of states \( C_i \) and formulated as

\[ F() = C_0, C_1, \ldots, C_k, \ldots, C_K \]  

where \( C_0 \) is the initial state; \( C_k \) is the final state; \( F() \) is the function of time warping that projects the time domain of one image to that of the other.

DP approach aims to find the \( F() \) function that optimizes the path from \( C_0 \) to \( C_k \). The minimal accumulated distance between the two images is calculated at the same time.

When searching for the optimal path, the main formula of DP is used at each step of the algorithm:

\[ D(\bar{x}_i, \bar{y}_j) = \min \begin{cases} D(\bar{x}_i, \bar{y}_{j-1}) + d(\bar{x}_i, \bar{y}_j) \\ D(\bar{x}_{i-1}, \bar{y}_j) + d(\bar{x}_i, \bar{y}_j) \\ D(\bar{x}_{i-1}, \bar{y}_{j-1}) + d(\bar{x}_i, \bar{y}_j) \end{cases} \]

where \( 0 \leq i \leq M, 0 \leq j \leq N \) and

\[ d(\bar{x}, \bar{y}) = \sum_{k=0}^{N_{SEC}} w^* (x_k - y_k)^2 \]

A weighted Euclidean metric is represented in (7), where \( x \) and \( y \) are vectors from the images compared, \( N_{SEC} \) is the dimension of feature vectors.

In order to get the distance at the final state it is necessary to calculate the distance matrix between the sequences of vectors. This is a very time consuming task demanding high computational power. In practice, two rows of the matrix are calculated, the succeeding calculated on the basis of the preceding. The process is iterative.

The distances between current frame frames and keyword templates are analyzed. The minimal distance is searched for and then compared with the threshold value.

2.5. Decision Thresholds

The threshold value introduced in the previous section may be changed manually. This is the way to tune hit / false alarm rate. The initial value of the threshold is obtained on the basis of the recognition results on the development corpus. If the threshold is increased, the number of false rejections reduces, but the false alarm rate grows up at the same time.

\[ T = \frac{1}{D} \ast 100\% \]  

In general, the threshold value is calculated as shown in formula 8. Distances between segments of the input speech streams and keyword templates are calculated. Then we choose the distance value that secures the desired hit / false alarm rate. This distance is referred to as \( D \) in the formula.

3. Evaluation

3.1. Evaluation Metrics

To evaluate a KWS system one needs to estimate several parameters:

- **Hits** (correct spotting) – number of correctly spotted keywords;
- **False Rejections** – number of missed keywords;
- **False Alarms** – number of non-keywords mistakenly accepted as keywords.

False alarm coefficient is also introduced in the evaluation. It is calculated as

\[ FA = \frac{N_{FA}}{N_{total} - N_{KW}} \]

where \( N_{fa} \) is the number of false alarms; \( N_{total} \) is the total number of words pronounced in the segment under consideration; \( N_{kw} \) is the number of keyword occurrences in the speech segment.

3.2. Results

Evaluation of the KWS system performance on the microphone databases was performed on the database containing audio files with technical texts read by male speakers with a microphone. The sound was digitized with a standard sound card. Keyword spotting was performed for three keywords (key phrases) ‘нечётный’, ‘программ’ and ‘капитолъ’ . The KWS was tested in two modes: with and without preliminary training. In the former case the system was tuned for each speaker voice. That means parameters of volume, timbre and pitch inherent to a speaker were set in the TTS module for generating synthesized templates. The results are presented in Table 1.

<table>
<thead>
<tr>
<th>Table 1: KWS in the Microphone Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preliminary Training</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>+</td>
</tr>
<tr>
<td>-</td>
</tr>
</tbody>
</table>

For testing KWS performance in telephone databases, speech files recorded in a telephone channel were analyzed.
Average length of files is 7 minutes. Keywords were different for each speaker (since keyword occurrence depends on the topic of conversation). The list of keywords for three speakers is presented in Table 2.

Table 2: Keywords Searched for in the Telephone Database

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>‘Ленинграде’, ‘Центре Речевых технологий’</td>
</tr>
<tr>
<td>3</td>
<td>‘День космонавтики’, ‘Эйзенхаузер’, ‘Аризоне’, ‘Рузвельт’</td>
</tr>
</tbody>
</table>

The results of KWS in the telephone database are presented in Table 3. No preliminary training was performed in this case.

Table 3: KWS in the Telephone Database

<table>
<thead>
<tr>
<th></th>
<th>Hits</th>
<th>False Alarms</th>
<th>False Rejections</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>78.5%</td>
<td>60%</td>
<td>21.5%</td>
</tr>
</tbody>
</table>

4. Conclusions

The system presented in this paper meets all requirements for KWS systems. It is able to work 24 hours a day unattended, performing search in huge. The number of input keywords is not limited either. However, the developers do not recommend entering more than 100 keywords into the search list simultaneously.

The set of tuning parameters makes the system flexible and easy to adapt to different speech databases.

One-pass method ensures fast search in the database. Search time for one keyword is less than 5% of the whole speech segment length. When more than one keyword is searched at a time, the search time grows nonlinearly, with time additions tending to get smaller for each additional keyword.

5. Future Directions

The major direction of future work is to add new synthesis voices to the system. That would make it more robust and boost the recognition results.

Another direction is to develop a multi-language TTS system in order to perform KWS not only for Russian, but also for other languages.

KWS systems find their implementation in telecommunication systems. Special normalization of the input signal is required for robust search in such systems.

Our evaluation has shown that the results depend on the parameters of the input signal. Best results were gained with manual correction of those parameters, i.e. volume, pitch and timbre (as was shown in Section 3.2). Gradation of these parameters resulted in another direction of work: automatic TTS adaptation to the input signal. The algorithm is supposed to analyze the input signal for obtaining optimal synthesis characteristics.

6. References