# WAKE-UP-WORD SPOTTING FOR MOBILE SYSTEMS

*A. Zehetner, M. Hagmüller, and F. Pernkopf*

Graz University of Technology
Signal Processing and Speech Communication Laboratory, Austria

## ABSTRACT

Wake-up-word (WUW) spotting for mobile devices has attracted much attention recently. The aim is to detect the occurrence of very few or only one personalized keyword in a continuous potentially noisy audio signal. The application in personal mobile devices is to activate the device or to trigger an alarm in hazardous situations by voice. In this paper, we present a low-resource approach and results for WUW spotting based on template matching using dynamic time warping and other measures. The recognition of the WUW is performed by a combination of distance measures based on a simple background noise level classification. For evaluation we recorded a WUW spotting database with three different background noise levels, four different speaker distances to the microphone, and ten different speakers. It consists of 480 keywords embedded in continuous audio data.

***Index Terms***— Wake-up-Word spotting, keyword spotting, dynamic time warping

## 1. INTRODUCTION

Single-phrase recognition systems can be roughly divided into three application perspectives, namely keyword spotting (KWS), Wake-up-word (WUW) detection, and spoken content retrieval (SCR). While the focus for each of these perspectives is different, they often rely on similar methods based on hidden Markov models (HMMs) and variants thereof.

KWS approaches aim to detect specific keywords within other words, sounds, and noises often without individually modeling the non-keywords. Some KWS approaches use individual HMM models for the keywords and filler or garbage models for non-keywords [1–3]. In [4], the detection of keywords in unconstrained speech without explicit modeling of non-keywords is addressed. They introduce a garbage/filler state at the beginning and end of each keyword HMM and an iterative Viterbi decoding is proposed to detect the *optimal* keyword boundary. Further, a widely used strategy is to search for keywords in phonetic lattices produced by large vocabulary continuous speech recognition (LVCSR) [5, 6]. These systems suffer from high error rates, especially when the speech is not clean [6, 7]. In [8] a hybrid two-stage system is proposed. An LVCSR system is used to produce word lattices. Then a sub-word approach is used to identify potential audio segments in the first stage. In the second stage, a more detailed search is performed verifying the candidate segments. Approaches for KWS relying on LVCSR require a considerable amount of speech resources. In [9], an approach for a limited amount of word-level transcription as annotated resource is proposed. In [10], acoustic KWS (i.e. keyword model is composed of phoneme models), spotting in word lattices generated by LVCSR, and an hybrid approach (i.e. searches in phoneme lattices generated by a phoneme recognizer) are compared. Other alternative approaches for KWS are based on large margin and kernel methods [11] or weighted finite-state transducers [12]. Recently, a fusion of several (probably) weak keyword detectors, each providing potentially complementary information, is performed [13]. So far, all these methods are computationally demanding and require transcriptions to adequately train the model parameters. To overcome this drawback of sufficient data, alternative techniques such as dynamic time warping (DTW) have been proposed [14]. DTW is based on matching a *template* to the test utterances. In the simplest case the recorded templates represented in feature domain can be used and at minimum *one* template keyword is sufficient. DTW optimally aligns the parametrized sequences of the template keyword and the acoustic input and determines the similarity between both. Recently, segmental DTW using Gaussian mixture models (GMMs) for representing each speech frame with a Gaussian posteriorgam has been proposed [15].

WUW speech recognition is related to KWS with the difference of detecting the token in an *alerting* context to wake up a permanently listening system [16]. Often speech recognition systems are activated by user interactions since continuously listening speech recognizers are insufficiently accurate. WUW recognition allows to activate these systems with speech commands.

This work has been supported by the European project DIRHA FP7-ICT-2011-7-288121 and the K-Project ASD. The K-Project ASD is funded in the context of COMET Competence Centers for Excellent Technologies by BMVIT, BMWFJ, Styrian Business Promotion Agency (SFG), the Province of Styria - Government of Styria and the Technology Agency of the City of Vienna (ZIT). The programme COMET is conducted by Austrian Research Promotion Agency (FFG).

SCR enables to search/browse audio data. Again the basic technology is LVCSR generating text transcripts from spoken audio. SCR is considered as information retrieval on ASR transcripts. So the task is to return content satisfying the user's request formulated by queries. However, the transcripts generally contain errors and audio is usually not structured into units such as paragraphs. KWS returns tokens matching the query phrase while in SCR the system returns *items that either treat the topic specified by a query of fit the description of that query* [17]. The goal in SCR is finding content that matches a query. Furthermore, the availability of information at query time can be different.

In this paper, we concentrate on WUW spotting in continuous audio using personal mobile devices. The aim is to trigger an alarm in emergency situations by speech input. We consider WUW spotting under the following constraints: speaker dependent spotting, use of individual personalized keyword, no transcription of the keyword, no training phase, low power and computing resources, and in general a language-independent system. With respect to these limitations, we propose template matching using DTW. The keyword is recorded by the naive user of the personal mobile device. This requires simple checks to ensure certain quality of the personalized keyword. The recognition is based on DTW combined with other distance measures depending on the background noise level. For evaluation we recorded a WUW spotting database with three different background noise levels, ten different speakers, and four different speaker distances to the microphone. In particular, the distance between speaker and microphone was 1m, 5m, and speaker was in an adjacent room to the recording device with either open or closed door. The background noise, i.e. television, was always at a distance of about 1m to the device. In total 480 keywords embedded in continuous audio data including speech and everyday sounds are recorded. The WUW algorithm evaluated on this challenging database has a recall of 59.6% and a precision of 99.7% where the focus in this application is on high precision (i.e more than 95%) and acceptable recall (i.e more that 50%). This means that about half of the WUW are detected correctly while almost all triggered detections are correct. This requirement is motivated by the associated costs of triggering too many false alarms in emergency applications. In case of an alarm the mobile system connects to a call center and an operator assists. A typical application are elderly people living alone.

The paper is organized as follows: Section 2 introduces the system for WUW spotting. In Section 3 the experimental setup, the recorded database and the results are presented. Section 4 concludes the paper.

## 2. WAKE-UP-WORD SPOTTING

The WUW spotting system is shown in Figure 1. Basically, only audio data blocks with an energy exceeding a threshold (block *energy detector*) are further analyzed. This helps to dramatically reduce the power-consumption.

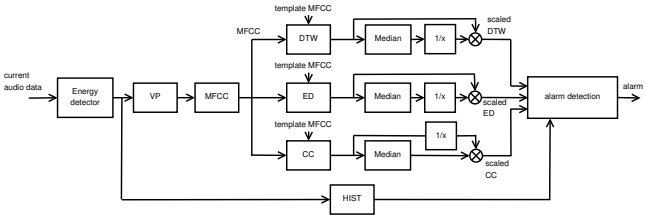


**Fig. 1**. WUW spotting system.

In case of sufficient energy in the audio data, voice pre-emphasis (VP) using a first-order FIR filter is performed to boost the energy in the high frequency components of the audio data. This flattens the spectrum of the audio signal. The pattern matching is performed in the cepstral domain. Therefore, the mel frequency cepstral coefficients (MFCCs) are determined for both the current audio signal and the reference recording. The $n^{\text{th}}$ signal frame of 32ms length and 16ms of overlap at a sampling frequency of 16kHz is represented by an MFCC feature vector $\mathbf{x}_n$. We noticed that the performance of WUW spotting improves when relying on more distance measures. For this reason, we introduce the euclidean distance (ED) and the cross-correlation (CC) between the MFCCs of the current audio signal and the keyword template in addition to DTW. This is further discussed in Section 2.2. All three distance measures are normed by their median filter values (Median), i.e. the scaling factor is the median of the past 20 distance measures. The final decision for detecting the WUW is based on the average of the distance measures[1] and the background noise detected in the audio frames. The background noise is classified based on the histogram of absolute amplitude values (block HIST in Figure 1).

### 2.1. Background Noise Classification (BNC)

The signal of each 10s audio frame (with 2s overlap) is classified into one of the following three categories: low-noise, medium-noise or high-noise background. Depending on the category of the current frame, a different threshold and combination of ED, CC, and DTW measures is used to detect the keyword. The categories are selected according to the distribution of the absolute amplitude values in the audio frame. The decision is based on the probability of absolute amplitude values falling in the first percentile $p_1$, i.e. we have *low-noise* if $p_1 \geq 0.80$, *medium-noise* if $0.80 > p_1 \geq 0.45$, and *high-noise* if $p_1 < 0.45$.

[1] Note that the CC measure is inverted.

### 2.2. Distance Measures and Alarm Detection

The DTW distance between the MFCCs $\mathbf{x}_i$ of the template consisting of $i = 1, \ldots, I$ frames and the MFCCs of the current audio signal is calculated for a block length of $I$ frames. It has been empirically observed that during reference recording hyper-articulation happens and the speaker talks slower than in general situations. Hence, we limit the analysis block length to the template length of $I$ frames. The hop-size for DTW computation is $0.1I$. The DTW values are normed by their median value. Similarly, the ED and CC distance measures are computed for the MFCCs of the currently processed signal blocked into matrices with $I$ MFCC frames with a shift of one frame (i.e. 16ms). To obtain the same number of distance values as DTW, blocks of $0.1I$ distances are averaged. Again, the ED and CC distances are scaled by their median value and CC distances are inverted. The CC distance can be easily determined in spectral domain by element-wise multiplication of the MFCCs of both the reference and the processed block and final summation.

Finally, for detecting the WUW the distance measures, i.e. ED, CC, and DTW of the 10s audio frame, are selected depending on the background noise class and averaged. If this averaged distance is below a threshold $T$, the WUW is present, otherwise not. We show results for different combinations of distance measures in Section 3.2.

### 2.3. Personalized WUW Recording

For emergency or other WUW applications it is important that naive, i.e. not trained, users are able to record individual personalized keywords. To ensure a good recognition performance, the reference, i.e. the template, is of essential importance. This makes simple checks necessary to guarantee a certain quality. Therefore, we advocate three measures to detect transient noise sources (e.g. slamming door), stationary noise sources (e.g. fan or traffic noise), and to ensure a rich phonetic content of the keyword. All measures have to meet the requirements, otherwise the keyword recording needs to be repeated. The intuition of these measures are listed in the following:

- Transient noise detection: We use the absolute value of the difference in signal energy between consecutive signal frames of 25ms and 5ms hop-size. These absolute values are averaged over five frames.
- Stationary noise detection: The keyword recording takes place in a pre-specified time window of 5s in a quiet environment. Within this window there has to be a significant signal energy difference between the very beginning and end of the window compared to the frames where the keyword is assumed.
- Rich phonetic content: Rich phonetic content in the keywords is supportive for good WUW spotting results. This means that keywords with just one vocal and no consonants like "*Ah*" should be rejected. This rejection is based on the amendatory zero-crossing rate [18] which relates to some extent the phonetic content of the keyword.

In case of sufficient quality of the personalized keyword, the user can listen again and confirm the reference recording.

## 3. EXPERIMENTS

The database is introduced and performance results for WUW spotting are presented based on the background noise classification and recording situations. The performance is measured via *recall* and *precision* where the aim is to achieve a precision of more than 95% at an acceptable recall of more than 50%. This means that about half of the WUW are detected correctly while almost all triggered detections are correct. This requirement is motivated by the costs of triggering too many false alarms in emergency applications caused by connecting to a call center. The best combination of the distance measures is determined by weighting the precision three times more important than the recall.

### 3.1. Database for WUW Spotting

This database consists of 10 different speakers – five female and five male – in three different background noise scenarios and four different distances to the microphone. In total there are 12 recordings per person, i.e. 3 background noise scenarios with 4 different distances. The distance of the recording device to the speaker is 1m, 5m, adjacent room with either open or closed door. The background noise source (i.e. television) is located in a distance of 1m to the microphone. For the low-noise scenario the television is switched off and only usual ambient noise from the environment (e.g. open window) is present. On each recording the speaker had to say the WUW, i.e. *Aktiviere Notruf!*, 4 times within 2 minutes. In total 480 keywords embedded in continuous audio data are recorded. Table 1 presents the mean and standard deviation of the a-posteriori signal-to-noise (SNR) ratio for the different background noise scenarios, i.e.

$$\mathit{SNR}[n] = 10 \log \frac{E\{|x[n]|^2\}}{E\{|w[n]|^2\}}, \tag{1}$$

where $E\{\cdot\}$ is the expectation operator, $x[n] = s[n] + w[n]$, $s[n]$ is the speech signal, and $w[n]$ is the noise signal.

### 3.2. Results: Combinations of ED, CC, and DTW

Table 2 presents the *precision* and *recall* results depending on the combination of the distance measures. Furthermore, results for the individual background noise scenarios (first three columns) and the whole database (last column) are shown. Using a combination of measures leads to better *precision*. The split of the evaluation into the background noise situations is necessary to find the best combination of

| measure combination | low-noise | | medium-noise | | high-noise | | all noise types | |
|---|---|---|---|---|---|---|---|---|
| | R | P | R | P | R | P | R | P |
| $ED$ | 79.4 | 100.0 | 58.8 | 91.3 | 38.8 | 86.1 | 59.0 | 93.7 |
| $DTW$ | 75.0 | 99.2 | 51.9 | 98.8 | 26.3 | 100.0 | 51.0 | 99.2 |
| $CC$ | 86.3 | 99.3 | 55.0 | 79.3 | 45.0 | 68.6 | 62.1 | 83.9 |
| $ED + CC$ | 85.0 | 100.0 | 57.5 | 92.0 | 42.5 | 90.7 | 61.7 | 95.2 |
| $ED + DTW$ | 76.3 | 99.2 | 53.1 | 98.8 | 30.0 | 100.0 | 53.1 | 99.2 |
| $CC + DTW$ | 78.8 | 99.2 | 55.6 | 98.9 | 30.6 | 98.0 | 55.0 | 98.9 |
| $ED + CC + DTW$ | 81.9 | 99.2 | 54.4 | 98.9 | 34.4 | 100.0 | 56.9 | 99.3 |

**Table 2**. Recall (R) and precision (P) results in [%] for different combinations of distance measures and background noise. The last column shows results for the whole database. The threshold of $T = 0.73$ has been empirically determined using development data.

| speaker distance | low-noise | | medium-noise | | high-noise | |
|---|---|---|---|---|---|---|
| | *SNR* | $\sigma$ | *SNR* | $\sigma$ | *SNR* | $\sigma$ |
| 1m | 18.4 | 8.6 | 8.2 | 7.0 | 6.9 | 4.8 |
| 5m | 21.5 | 6.7 | 6.3 | 6.6 | 4.3 | 2.7 |
| ORO | 15.7 | 8.0 | 5.2 | 7.2 | 3.5 | 3.6 |
| ORC | 4.3 | 3.5 | 1.5 | 4.5 | -0.3 | 4.7 |

**Table 1**. *SNR* and standard deviation in [dB] of different background noise scenarios. ORO is other room with open door, i.e. the speaker is in an adjacent room of the recording device with open door. ORC is other room with closed door.

the distance measures for the final evaluation using BNC in Section 2.1. If no BNC is used, only one combination can be used for all noise conditions. This is $ED + CC + DTW$ achieving high precision and acceptable recall performance.

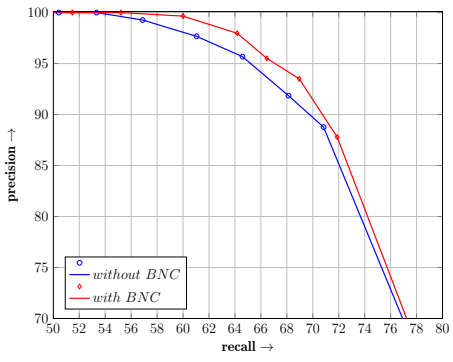


**Fig. 2**. Precision and recall curves of WUW spotting in [%] using different threshold $T$.

### 3.3. Results depending on SNR, Distance, and BNC

Based on the BNC a different combination of distance measures is used for the final decision. In Table 3(b) the results are split into background noise levels and speaker distances to the microphone. In (a) the best results without BNC using $ED + CC + DTW$ are summarized. It can be seen in (b) that using BNC improves the overall performance from $R = 56.9\%$ to $R = 59.6\%$ at a precision of $P = 99.3\%$ and $P = 99.7\%$, respectively. Furthermore, the recall drops dramatically from $100\%$ in low-noise and close speaker scenarios to $0\%$ in the challenging high-noise and speaker is in adjacent room at closed door situation. Note that the values in Figure 3(b) deviate from the values in Table 2 due to the BNC. The values in Table 2 are presented for the *true* background noise.

Moreover, Figure 2 shows the performance curves of the WUW spotting algorithm with and without BNC. The curves are obtained by varying the threshold $T$. BNC improves the recall while maintaining the precision. This is useful when the aim is to boost the recall at large precision values.

## 4. CONCLUSIONS

We developed a WUW spotting approach detecting only one personalized keyword in a continuous audio signal. The keyword is recorded by the naive user of the mobile system. The method combines several simple distance measures based on the background noise level estimate. The target application is mobile devices which limit the power-consumption and restrict the complexity of the algorithm. For evaluation we recorded a database with three different background noise levels and four different speaker distances to the microphone. In particular, the distance between speaker and microphone was 1m, 5m, and speaker was in an adjacent room than the device with either open or closed door. The overall performance on this challenging database is a recall of 59.6% and a precision of 99.7% where the focus in alerting applications is on high precision (i.e more than 95%) and acceptable recall (i.e more that 50%). This requirement is motivated by the costs of triggering too many false alarms in emergency applications where a connection to a call center is established.

## REFERENCES

[1] Y. Benayed, D. Fohr, J-P Haton, and G. Chollet, "Confidence measures for keyword spotting using support vector machines," in *ICASSP*, 2003, pp. 588–591.

[2] H. Ketabdar, J. Vepa, S. Bengio, and H. Bourlard,

(a)

| | **low-noise** $ED + CC + DTW, T = 0.73$ | | | | **medium-noise** $ED + CC + DTW, T = 0.73$ | | | | **high-noise** $ED + CC + DTW, T = 0.73$ | | | | **all noise types** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1m | 5m | ORO | ORC | 1m | 5m | ORO | ORC | 1m | 5m | ORO | ORC | |
| **R** | 100.0 | 85.0 | 90.0 | 52.5 | 85.0 | 70.0 | 55.0 | 7.5 | 65.0 | 32.5 | 40.0 | 0.0 | |
| **P** | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 75.0 | 100.0 | 100.0 | 100.0 | 100.0 | |
| **R** | 81.9 | | | | 54.4 | | | | 34.4 | | | | 56.9 |
| **P** | 99.2 | | | | 98.9 | | | | 100.0 | | | | 99.3 |

(b)

| | **low-noise** $ED + CC, T = 0.77$ | | | | **medium-noise** $CC + DTW, T = 0.73$ | | | | **high-noise** $ED + CC + DTW, T = 0.73$ | | | | **all noise types** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1m | 5m | ORO | ORC | 1m | 5m | ORO | ORC | 1m | 5m | ORO | ORC | |
| **R** | 100.0 | 90.0 | 97.5 | 67.5 | 85.0 | 70.0 | 57.5 | 10.0 | 65.0 | 32.5 | 40.0 | 0.0 | |
| **P** | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 80.0 | 100.0 | 100.0 | 100.0 | 100.0 | |
| **R** | 88.8 | | | | 55.6 | | | | 34.4 | | | | 59.6 |
| **P** | 100.0 | | | | 98.9 | | | | 100.0 | | | | 99.7 |

**Table 3**. Recall (R) and precision (P) results in [%] of the WUW algorithm. (a) Without using BNC $ED + CC + DTW$ and $T = 0.73$ are used for all background noise scenarios. (b) With BNC the best threshold and distance measure combination for each background noise scenario is selected. ORO is other room with open door, i.e. the speaker is in adjacent room of the recording device with open door. ORC is other room with closed door. The last column shows results for the whole database.


"Posterior based keyword spotting with a priori thresholds," in *Interspeech*, 2006.

[3] R.C. Rose and D.B. Paul, "A hidden Markov model based keyword recognition system," in *ICASSP*, 1990, pp. 129–132.

[4] M.-C. Silaghi and H. Bourlard, "Iterative posterior-based keyword spotting without filler-models: Iterative viterbi decoding and one-pass approach," Tech. Rep., 2000.

[5] P. Yu and F. Seide, "Fast two-stage vocabulary-independent search in spontaneous speech," in *ICASSP*, 2005, pp. 481–484.

[6] F. Seide, P. Yu, C. Ma, and E. Chang, "Vocabulary-independent search in spontaneous speech," in *ICASSP*, 2004, pp. 253–256.

[7] R. Ordelman, F. de Jong, and M. Larson, "Enhanced multimedia content access and exploitation using semantic speech retrieval," in *IEEE International Conference on Semantic Computing*, 2009, pp. 521–528.

[8] A. Norouzian and R. Rose, "An approach for efficient open vocabulary spoken term detection," *Speech Communication*, vol. 57, pp. 50–62, 2014.

[9] A. Garcia and H. Gish, "Keyword spotting of arbitrary words using minimal speech resources," in *ICASSP*, 2006, pp. 949–952.

[10] I. Szöke, P. Schwarz, P. Matejka, and M. Karafiat, "Comparison of keyword spotting approaches for informal continuous speech," in *Eurospeech*, 2005.

[11] J. Keshet, D. Grangier, and S. Bengio, "Discriminative keyword spotting," *Speech Communication*, vol. 51, pp. 317–329, 2009.

[12] Y. Guo, Z. Zhang, T. Li, J. Pan, and Y. Yan, "Improved keyword spotting system in weighted finite-state transducer framework," *Journal of Computational Information Systems*, vol. 9, no. 12, pp. 4947–4954, 2013.

[13] A. Abad and R.F. Astudillo, "The l2f spoken web search system for mediaeval 2012," in *Mediaeval 2012 Workshop*, 2012.

[14] J. Bridle, "An efficient elastic-template method for detecting given words in running speech," in *Britisch Acoustic Society Meeting*, 1973.

[15] Y. Zhang and J.R. Glass, "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams," in *ASRU*, 2009, pp. 398–403.

[16] V.Z. Këpuska and T.B. Klein, "A novel wake-up-word speech recognition system, wake-up-word recognition task, technology and evaluation," *Nonlinear Analysis*, vol. 71, pp. 2772–2789, 2009.

[17] M. Larson and G.J.F. Jones, "Spoken content retrieval: A survey of techniques and technologies," *Foundations and Trends in Information Retrieval*, vol. 5, no. 4-5, pp. 235–422, 2012.

[18] H. Zhao, L. Zhao, K. Zhao, and G. Wang, "Voice activity detection based on distance entropy in noisy environment," in *Int. Conf on INC, IMS and IDC*, 2009, pp. 1364–1367.