

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/354095669>

Voice activation using prosodic features

Conference Paper · October 2004

DOI: 10.21437/Interspeech.2004-753

CITATIONS

4

READS

37

4 authors, including:



Matthias Wolff

Brandenburg University of Technology Cottbus - Senftenberg

211 PUBLICATIONS 879 CITATIONS

SEE PROFILE



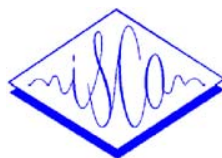
Matthias Eichner

75 PUBLICATIONS 346 CITATIONS

SEE PROFILE

All content following this page was uploaded by Matthias Eichner on 02 September 2021.

The user has requested enhancement of the downloaded file.



VOICE ACTIVATION USING PROSODIC FEATURES

Marco Kühne, Matthias Wolff, Matthias Eichner and Rüdiger Hoffmann

Dresden University of Technology
Laboratory of Acoustics and Speech Communication
D-01062 Dresden, Germany

Abstract

In this paper we propose a voice activation method based on prosodic keyword verification. In current voice activation systems features like the fundamental frequency contour have not been considered so far. Normally a continuous listening word spotter is used to detect a certain predefined keyword. We conducted an experiment which shows that people emphasize this keyword when they address a recognizer. To capture the prosodic information we trained an HMM on the fundamental frequency and energy contour of the keyword. The prosodic model is used to verify the keyword hypotheses of a phonetic recognizer. We investigated the performance of the prosodic model to distinguish between the same keyword spoken in command and non-command phrases. The introduction of the prosodic information significantly reduced the false alarm rate whereas the detection rate was only slightly degraded.

1. Introduction

One crucial point for the practical use of speech recognizers is the activation of the system. Many current systems require a “push-to-talk” activation. However, for practical applications such a manual activation is unsuitable. A voice activation of the recognizer is often more ergonomic as it enables a complete hands- and eyes-free speech interface. Voice activation systems usually consist of a continuous listening word spotter which detects a certain predefined keyword or key phrase [1]. They use a word HMM for the keyword and represent non-speech and non-keyword segments by garbage and filler models. The greatest challenge for a voice activation system is the robust rejection of out-of-vocabulary (OOV) words in order to avoid false alarms [2]. But even in the case of an ideal word spotter false alarms occur if the keyword is uttered in another context than a command. For example an activation keyword like “computer” will most likely occur outside command phrases, too. The common solution to that problem is choosing exotic, but not necessarily user-friendly, keywords which occur rarely in normal speech. However, people also emphasize a keyword in some way when they address a recognizer. Thus we can use prosodic features like the fundamental frequency, the intensity or the word or syllable duration to determine whether the keyword was uttered as part of a voice command or not. In current voice activation systems the fundamental frequency and the intensity have not been considered as prosodic features so far. However, Yamashita and Mizoguchi showed in [3] that the performance of a keyword spotter benefits from additional prosodic knowledge. They calculated the prosodic similar-

ity between a keyword reference pattern and input speech through a Dynamic Time Warping approach. However, this approach is speaker dependent. For a speaker independent system selecting only one template as prosodic model is inadequate. Therefore we propose to model the prosodic features of the keyword by an HMM. The primary goal of our work is a more robust voice activation without using exotic keywords through the combination of phonetic recognition and prosodic verification.

2. The prosodic model

In [4] Ljolje and Fallside showed the capability of HMMs to model the essential features of isolated prosodic patterns in terms of fundamental frequency, timing and intensity. We use the fundamental frequency F_0 , the RMS energy and their first time derivatives as features to describe the prosodic patterns of keywords intended as a command. The calculation of raw F_0 and RMS features was done by the widely used ESPS tool `get_f0`. Observation vectors were calculated every 10 milliseconds. Voiced segments shorter than 40 milliseconds were rejected and reset to zero. Unvoiced segments were interpolated by a spline function to achieve a continuous F_0 contour. For a speaker independent processing RMS and F_0 contours were normalized, smoothed by an average filter and converted to a log-scale. Figure 1 shows average F_0 contours of the keyword “computer” for 30 Speakers. Contours of keywords intended as a command are printed in gray, contours of keywords occurring in normal speech are printed in black. The speech samples were recorded during speech controlled slide presentations (see section 4.2). The speech melody significantly differs between command and normal speech. Thus the fundamental frequency can be used as a discriminative feature. To model the prosodic features we trained a seven state left-to-right *prosodic* HMM

$$PHMM = \{V, E, \{\mathcal{N}\}, \nu^{(V)}, \pi^{(E)}\} \quad (1)$$

consisting of a node set V and a directed edge set $E \subseteq V \times V$. Through the map $\nu^{(V)} : V \rightarrow \{\mathcal{N}\}$ a single 4-dimensional Gaussian probability density function $\mathcal{N}_i(\underline{\mu}_i, \Sigma_i) \in \{\mathcal{N}\}$ with full covariance matrix Σ_k is assigned to each node. Each edge carries a transition probability $\pi^{(E)} : E \rightarrow \mathbf{R}^{(0,1]}$.

We performed a Viterbi training with 138 recorded samples of 3 female and 3 male German speakers. All training samples were visually inspected for obvious failures of the automatic F_0 extraction. Erroneous samples were removed. The Gaussians were initialized by distributing the feature vectors uniformly over the model states.

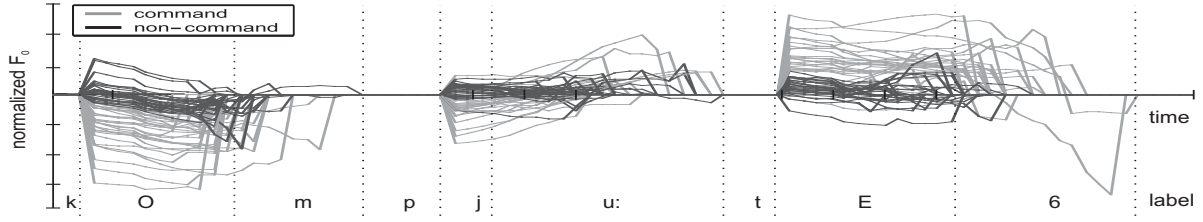


Figure 1: F_0 contours of the keyword “computer” taken from command and non-command phrases

During the classification procedure we build a linear graph $\mathcal{X}$ from the observation feature vector sequence $\underline{X}$ which assigns exactly one feature vector to each node of the graph. Then the graph $\mathcal{X}$ is matched on the trained model PHMM using a Viterbi search:

$$\mathcal{U}^* = \underset{\mathcal{U} \subseteq \text{PHMM} \times \mathcal{X}}{\operatorname{argmax}} \left[\sum_{i=1}^{|\mathcal{U}|} D_{LL}(\underline{x}(u_i) | \mathcal{N}(u_i)) \right] \quad (2)$$

where $\mathcal{U}^*$ denotes the path through PHMM which maximizes the sum of the log-likelihoods of the feature vector $\underline{x}$ given the Gaussians $\mathcal{N}(v_k)$:

$$D_{LL}(\underline{x} | \mathcal{N}(v_k)) = -(\underline{x} - \underline{\mu}_k)^T \Sigma_k^{-1} (\underline{x} - \underline{\mu}_k) - \ln |\Sigma_k| \quad (3)$$

We write $LL^*(\underline{X} | \text{PHMM})$ for the log-likelihood sum along the best path $\mathcal{U}^*$. To incorporate state duration information into the PHMM we implemented a heuristic method based on normalized state duration histograms [8]. During the training stage normalized state duration histograms are created for every state. These histograms are used to estimate state duration probability density functions $p(\tau_i)$ which are considered to be Gaussian with 2 to 5 mixtures. In the recognition stage the normalized state durations τ_i are calculated after the Viterbi segmentation is done. Then, the log-likelihood value LL^* of the optimal state sequence is incremented by the log of the state duration probability $p(\tau_i)$:

$$PS = LL^*(\underline{X} | \text{PHMM}) + \eta \sum_{j=1}^N \log [p_j(\tau_j)] \quad (4)$$

where η is a scaling factor. We use this modified log-likelihood sum as a prosodic confidence measure PS to determine how similar the observations $\underline{X}$ and the trained model PHMM are.

3. Command recognition with voice activation

For our experiments we equipped the SMG [7] based UASR [5] recognizer with the prosodic model described above. The acoustic models were trained with approximately 30 hours of machine labeled spontaneous speech taken from the German VERBMOBIL database [6]. The recognizer was configured for command phrases using a small lexicon and an explicit command grammar. One command phrase consists of

the keyword “computer” followed by a command. After the acoustic recognition the PHMM was utilized to verify the keyword hypotheses of the recognizer.

The command phrase recognizer applies two confidence measures to reject OOV utterances. Both use an unconstrained phoneme recognition as reference. The command recognition and the reference recognition employ identical SMG monophone models. The first confidence measure is calculated at the feature vector level and compares the number of matching phoneme symbols between command and reference recognition. The higher the RPMS the better is the *phonetic* confidence. We called this the relative phonetic match score (RPMS):

$$RPMS = \frac{\text{number of matched phoneme labels}}{\text{number of all phoneme labels}} \quad (5)$$

The second confidence measure combines the acoustic and prosodic log-likelihood scores. It is called relative acoustic score difference (RASD):

$$RASD = \left| \frac{(AS_{cmd} - AS_{ref}) - w(|PS| - |\overline{PS}|)}{AS_{cmd}} \right| \quad (6)$$

AS_{cmd} denotes the log-likelihood of the command recognition, AS_{ref} represents the log-likelihood of the reference recognition and $w(|PS| - |\overline{PS}|)$ is a prosodic penalty from the PHMM which can be scaled by the factor w . $|\overline{PS}|$ is the average prosodic score of all keywords [3]. The smaller the RASD the higher is the *acoustic* and *prosodic* confidence of the recognition result. A weight $w = 0$ gives a (baseline) system BL not using any prosodic information in contrast to a system VA with $w > 0$ which applies the proposed prosodic keyword verification.

We tested two rejection strategies. In the first one (OP) predefined thresholds α and β are used which mark a fix operating point $OP(\alpha, \beta)$. A recognition result will be accepted if $RASD \leq \alpha$ and $RPMS \geq \beta$.

The second strategy (CMC) treats the rejection process as a confidence measure classification task. We assume that the distribution of the confidence measures for the classes $k \in \{cmd, oov\}$ can be described by a 2-dimensional Gaussian with mean vector $\vec{\mu}_k$ and covariance matrix Σ_k . The vector $\vec{c} = (RASD \quad RPMS)^T$ represents both confidence values of a recognition result. Based on the Mahalanobis-distance

$$d_k(\vec{c}) = (\vec{c} - \vec{\mu}_k)^T \Sigma_k^{-1} (\vec{c} - \vec{\mu}_k) \quad (7)$$

speech input will be rejected if $d_{cmd}(\vec{c}) > d_{oov}(\vec{c})$.

4. Experiments

The voice activation was evaluated on a speech controlled slide presentation given by 24 male and 6 female German speakers. All speech material was recorded at 16 kHz sampling frequency, 16 bit quantization with a headset microphone in a laboratory environment. The speech data contain 344 phrases with the keyword “computer” followed by a command to control the presentation (e.g. “next slide”) and 193 phrases that start with the keyword but were part of the talk and not intended as commands. The remaining data were automatically segmented into 3630 OOV utterances of approximately the length of a command phrase. The whole database comprises 0.69 hours of speech. The speakers were *not* told to use a special speaking style to address the recognizer. The vocabulary consisted of 7 words, an explicit grammar defines four different commands on this vocabulary.

The experiments were divided into three different tasks. In the first task we investigate the influence of the prosodic weight w (see equation 6) on the performance of the command recognizer. The second task evaluates the ability of the prosodic model to distinguish between keywords in commands and normal speech. It also compares the performance of the PHMM with the results of human test subjects in two listening tests. The third task examines the performance of the command recognizer, particularly the rejection of OOV utterances.

4.1. Task 1: Choosing the prosodic weight

Figure 2 shows the command detection rate (DR) and the false alarm rate (FAR) for different prosodic weights w in a fix operating point $OP_1(0.09,0.25)$. The baseline system BL achieves a detection rate of 95.1 % at a false alarm rate of 28.5 % at the test set described above. With increasing prosodic weight the false alarm rate drops significantly whereas the detection rate is only slightly degraded. For the following experiments we use the prosodic weight $w = 0.2$. Using this weight the VA system achieves 90.1 % detection rate at 4.5 % false alarm rate.

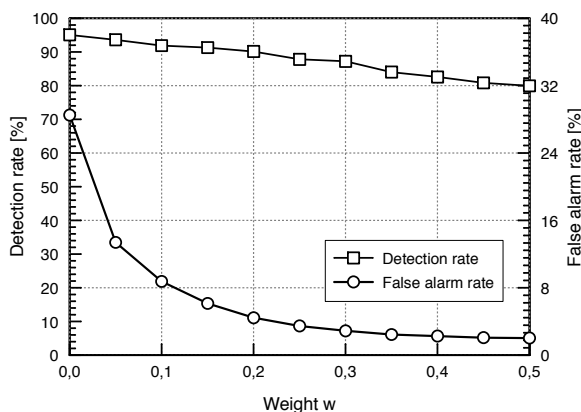


Figure 2: Effect of the prosodic weight w

4.2. Task 2: PHMM vs. human performance

In this experiment 50 isolated instances of the keyword “computer” taken from the test set were automatically classified as “command” or “normal speech” using the prosodic model described in section 2. Additionally we performed two listening tests with 32 test persons doing the same classification task. As the listening samples contained only the keyword the subjects were forced to make their decision purely based on prosodic information. The keyword instances were uttered by 4 male and 4 female speakers. In the first test (T1) the samples from all speakers were randomly arranged. In the second test (T2) the samples were grouped by the eight different speakers. To provide the subjects with further information on the speaking style a short sentence of the current speaker was presented in T2. Table 1 compares the average detection and false alarm rates for the test listeners with the automatic classification. In T1 the human performance is only slightly better than the automatic classification which indicates a quite good performance of the PHMM. The results of the test T2 reveal a good ability of humans to adapt to different speaking styles when they were supported with more speaker dependent information. This suggests to apply speaker adaptation techniques on the PHMM in the future.

	Humans		PHMM Model
	T1	T2	
Detection rate [%]	81.1	88.3	74.2
False alarm rate [%]	29.6	19.3	25.5

Table 1: Results of the listening tests and the automatic classification PHMM

4.3. Task 3: Command recognition

To benchmark the prosodic verification we compare a baseline system BL without prosodic information with a system VA using the proposed prosodic keyword verification. Figure 3 shows the performance of both systems for several operating points of the recognizer. For instance, at 10 % false rejection rate, the proposed system VA resulted in 4 % false alarm rate. This clearly outperforms the baseline

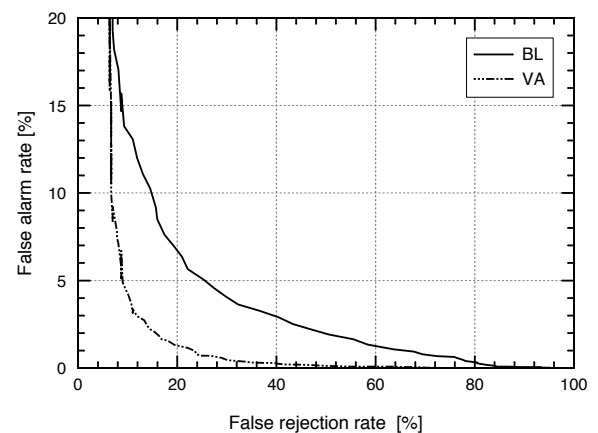


Figure 3: Comparison of the baseline system BL and the proposed voice activation VA ($w = 0.2$)

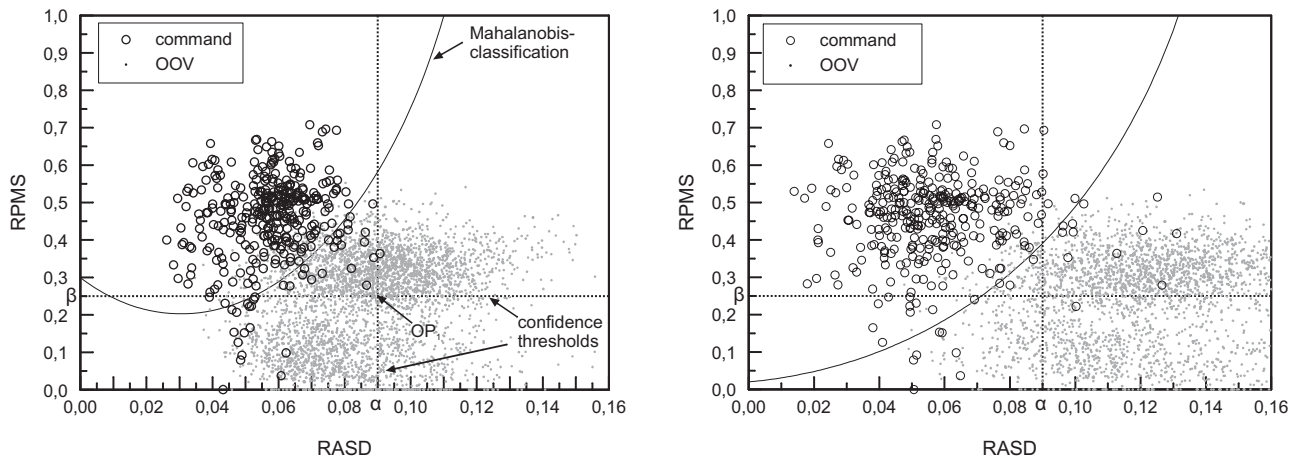


Figure 4: Distribution of the confidence measures RASD and RPMS for a prosodic weight $w = 0$ (BL, left) and $w = 0.2$ (VA, right).

system BL which achieved a 14 % false alarm rate at the same 10 % false rejection rate. The equal error rate of the baseline system BL was reduced from 13.1 % to 6.9 % in the case of the VA system.

Figure 4 shows the distribution of the confidence measures RASD and RPMS with and without prosodic keyword verification. Further the Mahalanobis discriminant function is shown in contrast to the two rejection thresholds which mark the operating point OP_1 . This Mahalanobis discriminant function was used to evaluate the second rejection strategy. Compared to the baseline system BL which achieved a 89.0 % detection rate at 6.1 % false alarm rate, both rates were improved by the additional prosodic verification. For a weight $w = 0.2$ the VA system resulted in a 92.2 % detection rate at 2.7 % false alarm rate. Based on a representative training set suitable discriminant functions can be estimated to reject OOV utterances.

Table 2 summarizes the experimental results of the systems BL and VA for both rejection schemes.

	BL _{OP}	VA _{OP}	BL _{CMC}	VA _{CMC}
DR [%]	85.2	91.4	89.0	92.2
FAR [%]	13.1	6.9	6.1	2.7
FRR [%]	13.1	6.9	10.3	7.7

Table 2: Detection rate, false alarm and false rejection rate (DR, FAR, FRR) for both systems BL and VA depending on the used rejection strategy.

5. Conclusions

Experiments show that humans emphasize the keyword of a voice activated speech recognizer in some way. We proposed a voice activation system based on prosodic keyword verification. To capture the prosodic information we trained a Hidden Markov Model using fundamental frequency and energy

based prosodic features of the keyword. We showed that a voice activation using prosodic verification significantly outperforms a conventional recognizer without prosodic information. In our future work we will implement a pause detection tailored to voice activation and investigate methods of speaker adaption for an enhanced prosodic modeling.

6. References

- [1] Sahar E. Bou-Ghazale, and Ayman O. Asadi. Hands-free Voice Activation of Personal Communication Devices. In *Proc. ICASSP*, Istanbul, 2000.
- [2] Iso-Sipilä, J. and Laurila, K., and Hariharan R., and Viiki O. Hands-free Voice Activation in Noisy Car Environment. In *Proc. Eurospeech*, Budapest, 1999.
- [3] Yamashita, Y. and Mizoguchi R. Keyword Spotting Using F0 Contour Matching. In *Proc. Eurospeech*, Rhodes, 1997.
- [4] Ljolje, A. and Fallside, F. Recognition of isolated prosodic patterns using Hidden Markov Models. In *Computer Speech and Language*, pp. 27–33, 1987.
- [5] Eichner, M. and Wolff, M. and Hoffmann R., A Unified Approach for Speech Synthesis and Speech Recognition Using Stochastic Markov Graphs. In *Proc. ICSLP*, Beijing, 2000
- [6] Wahlster, W., editor. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer-Verlag Berlin Heidelberg, 2000.
- [7] Wolfertstetter, F. and Ruske, G. Structured Markov models for speech recognition. In *Proc. ICASSP*, pp. 544–547, 1995.
- [8] L. R. Rabiner, A tutorial on Hidden Markov models and selected applications in speech recognition. In *Proc. IEEE*, Vol. 77, pp. 257–286, 1989.