

# HANDS-FREE VOICE ACTIVATION IN NOISY CAR ENVIRONMENT

*J. Iso-Sipilä, K. Laurila, R. Hariharan, O. Viikki*

Nokia Research Center, Speech and Audio Systems Laboratory, Tampere, Finland

Email: {juha.iso-sipila,kari.laurila,ramalingam.hariharan,olli.viikki}@research.nokia.com

## ABSTRACT

In this paper, we propose an algorithm that enables hands-free and eyes-free voice activated dialing in noisy car environment. Noise robustness is achieved by duration modeling and feature vector normalization techniques. High acceptance of valid command words and high rejection of out-of-vocabulary input is obtained by combination of command word partitioning into proper subword models and a user interface related multi-level approach which utilizes users' custom to repeat unrecognized commands. Experimental simulations show that both in, speaker-independent and speaker-dependent case, we are capable of achieving 99% acceptance of valid command words with less than one false alarm in 25 hours of conversational speech, music and car noise. In addition, real-time car environment tests with novice users show the validity of the simulation results.

Keywords: Keyword spotting, noise robust speech recognition, utterance verification

## 1. INTRODUCTION

The final goal in the development of automatic speech recognition algorithms is to enable human to human like free conversation with machines. Despite huge improvements achieved during recent years, speech recognition community is still far from this goal, measured either in recognition accuracy or in years.

Currently many recognizers require users to express themselves with some non-speech modality when they start speaking. In human-machine dialogues, users are often given only some seconds to speak, otherwise the system takes over. Systems which continuously listen to the user are rare. One approach towards continuous listening is to utilize word spotting techniques [1,2]. This approach provides user a hands-free and eyes-free way to initiate a dialogue with a machine by speaking a predetermined command or phrase. That is, word spotting does not enable free conversation, but allows the user to start a dialogue with speech. We call this approach voice activation.

Compared to push-to-talk, voice activation brings clear advantages, particularly in hands-busy situations. In car environment, voice activation enables user to use both hands for steering and to keep eyes on the road. Especially with professional mobile radio (PMR), users may quite often have both their hands busy. For example, users may wear gloves or they may have dirty hands etc., which makes calling using a keypad rather awkward. This is also usually the case with people working in warehouse inventory. In general, voice

activation enables remote control of devices with speech, no touching is required.

Why isn't voice activation more common then? At least partly one has to blame technology. One of the biggest problems in automatic speech recognition is the poor rejection of out-of-vocabulary (OOV) words. For a recognizer it is difficult to decide if an utterance spoken by the user belongs to the vocabulary or not. In voice activation this decision is crucial. It is required that the system should not respond to general speech, radio program or car noise and still have a high acceptance rate for the correct command words.

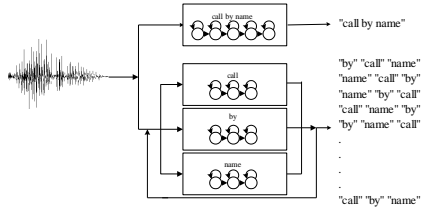
Usually a garbage or filler model is utilized for the detection of OOV words [1,2]. In [3] a special garbage model was introduced, where the probability of the garbage model was calculated from the word model probabilities. Recent work in [4] uses inter-frame correlation for utterance verification.

In this paper, we propose novel subword modeling, multi-level recognition and utterance detection approaches tailored to voice activation. By means of speaker-independent and speaker-dependent experiments we show that the proposed approach significantly outperforms the baseline word spotting scheme.

## 2. SUBWORD HMM APPROACH

A high rejection rate of OOV words is a key requirement of voice activation systems. The user finds it very irritating if a system activates from background speech, noise, or other non-keyword speech input. In order to make voice activation applications attractive to the users, the false alarm rate should be as low as possible. On the other hand, the acceptance rate of valid activation phrases should be as high as possible. To meet both of these high performance objectives, one must develop special techniques focusing on OOV word rejection.

An effective approach to improve the rejection performance is to use subword modeling for characterizing activation phrases. Conventional left-to-right whole-word HMM structure can provide only one state order. When having subword models this constrained structure can be relaxed as illustrated in Figure 1. If subword models are recognized with a loop grammar, i.e. in any order, we are capable of introducing an additional degree of freedom. If correct utterance has truly been spoken, it is likely that the subword HMMs are recognized in a proper order, whereas in the case of out-of-vocabulary utterance, the subword models are recognized in an arbitrary order.



**Figure 1.** Whole-phrase and loop grammar subword modeling approaches to voice activation.

When subword HMMs are recognized in a proper order, the validity of the recognized utterance can further be tested using confidence measures, such as log-likelihood ratio test. Using the subword approach, the rejection performance is proportional to  $N^V$  where  $N$  is the total number of subword models. When aiming at extremely high OOV word rejection rate without substantially reducing the acceptance rate of valid activation phrases, the grammar validity test is essential as OOV words occasionally tend to produce high confidence values, which lead to false activations.

## 2.1 Subword Modeling Alternatives

Since phonemes are one of the basic units of speech, it is straightforward to apply them as subword HMMs. However, a typical activation phrase consists of several phonemes, e.g. in Figure 1 there are 10 phonemes corresponding to  $10^{10}$  possible combinations. As the number of subword models increases, it becomes more difficult to recognize subword models in a proper order. To solve this problem, one can accept invalid combinations which are likely to occur when correct activation phrase is spoken. To avoid the use of complex post-processing techniques, we propose here an alternative method for subword modeling for voice activation systems.

One can directly construct the subword models from the whole-word HMM which describes the activation phrase. Subword models are then obtained by dividing the whole-word HMM into  $N$  subword units as illustrated in Figure 1. In practice, the number of subword models should be  $N \geq 3$  so that a sufficient OOV word rejection rate can be achieved. With this technique, we can use only as many subword models as is needed. When creating the subword models, one has to pay special attention to the splitting process to prevent the subword models to be too similar. If two or more subword HMMs are too close to each other, it is difficult to recognize these models in a proper order even in the case of valid activation phrase.

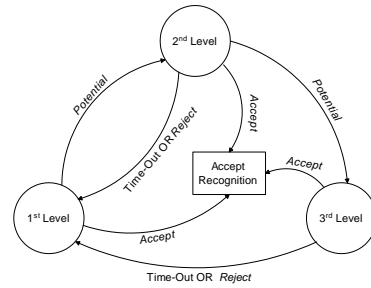
## 3. MULTI-LEVEL APPROACH

It is natural for humans to repeat unrecognized commands, i.e. when the user has spoken a command with no response from the system. This behavior is exploited in our voice activation system in order to further enhance the performance.

In order to guarantee high OOV word rejection rate, it is necessary to have high confidence thresholds. However, this leads to low acceptance rate of correct utterances. To solve this problem, one can continuously monitor whether utterances have confidences close to the required values. If such confidences are observed, the user might have said the proper activation phrase, and is likely to repeat it within a few seconds.

In our system, each result candidate is classified as one of the three classes: *accept*, *reject* or *potential*. *Accept* means that the confidence of the utterance is high enough and the recognition can be accepted. *Reject* means that the confidence is too low and the recognition is rejected. *Potential* means that the confidence is neither too low nor high enough for the two former classes. If the result candidate is classified as *potential*, the recognizer changes state for  $T$  seconds and then returns to normal operation mode.

Our multi-level approach consists of  $M$  states. Each higher state has lower confidence thresholds than the previous lower state. This means that each time a state change occurs to a higher state, the recognizer accepts the result candidate more easily. Figure 2 shows a voice activation state machine with  $M = 3$ . Here, the 1<sup>st</sup> level corresponds to the normal operation mode of the recognizer. A state change occurs when a result candidate is *potential*. The state is changed back to the 1<sup>st</sup> level when a time-out or *reject* occurs. The result candidate can be accepted in each level.



**Figure 2.** Multi-level voice activation approach with three levels.

Lower thresholds in higher levels increase the probability of false alarms. To avoid this, the successive repetitions are compared in order to ensure that they are the same.

Whenever a result candidate is obtained in a higher level  $m$  ( $m > 1$ ), the utterance in level  $m$  is compared to the utterance in level  $m-1$ . The comparison is done at the feature vector level and it uses a Viterbi-algorithm based time alignment scheme. The distance between the two repetitions is then used in determining whether the repetitions are the same phrase or not. If they are close enough and the recognition is classified as *accept*, the result candidate at level  $m$  is accepted. If the distance is too high, the result candidate is rejected and the level is changed to the 1<sup>st</sup> level.

Sometimes OOV input makes the recognizer change the level. It is very unlikely that two such inputs that produce high confidences, occur within  $T$  seconds. Furthermore, the repetitions are still compared which reduces the possibility of OOV input to activate the recognizer.

## 4. CONFIDENCE CALCULATION

A confidence measure is used to determine whether a recognition result is correct or not. This confidence is based on the difference between the word model and the garbage model log-likelihood scores [5]. The confidences of each subword model are normalized with the length of the recognized subword. Hence,  $N$  confidences are obtained for each recognized activation phrase. These confidences are compared against the pre-defined confidence thresholds.

## 5. UTTERANCE DETECTION

It is assumed that the voice activation phrase is not embedded in general speech. This fact is exploited to enhance the OOV word rejection capability in our approach. We require that the command word candidate is preceded and followed by periods of silence. Even though there are very few false alarms using the multi-level scheme, this added layer of protection tries to ensure that the valid command words embedded in general speech, if any, do not trigger the activation stage.

The proposed algorithm uses energies of  $S$  subbands (frequency bands) to detect the end-of-utterance. Subbands are created by decimation of Mel frequency band energies produced by a typical MFCC front-end. The algorithm is based on finding the time instant when each subband energy falls below an adaptive threshold, in other words a pause is detected after something has been uttered. A further requirement is that the length of the pause period must be large enough to avoid the detection of pauses between words. If many enough individual subbands satisfy the end-of-utterance detection criterion separately, then the end-of-utterance of the spoken command is detected.

Computation of the threshold for comparison with the subband energy values requires the estimation of global maximum  $P_{max}$  and minimum  $P_{min}$  energy values in each subband. These values are updated in each frame based on the subband energy values in the current frame. The threshold  $P_{thr}$  is then computed using the following equation:

$$P_{thr} = P_{min} + k \cdot (P_{max} - P_{min}), \quad (1)$$

where  $k$  ( $0 < k < 1$ ) is a constant. The global maximum and minimum energy values are also adapted to compensate for any change in environments during the voice activation stage. The end-of-utterance for the subband is detected when the energy value remains below the threshold for a required amount of frames.

The pause period corresponding to the beginning of the activation phrase is also determined after the detection of the end-of-utterance. When end-of-utterance is detected, it is still verified that there is a long enough pause preceding the utterance. These two criteria should be met before further processing.

## 6. EXPERIMENTS

Both Speaker-Independent (SI) and Speaker-Dependent (SD) experiments were carried out to evaluate the performance of the proposed algorithms. The training data of SI models consisted of 300 speakers each uttering the activation phrase once. Three noisy versions of each training utterance were obtained by artificially adding car noise to the utterances. Multi-environment state duration constrained [6] HMMs were estimated with Maximum Likelihood training. An MFCC front-end with recursive feature vector normalization [7] was used.

The speaker-dependent training procedure utilized three samples of each activation phrase. Artificial noise mixing and multi-environment training was used in SD case as well.

The test database consisted of 15 Finnish speakers saying two Finnish activation phrases. Each activation phrase was uttered three times per recording with a 2-3 second pause between

repetitions. There were 40 recordings per activation phrase and speaker. The overall amount of recordings in the database was 1200.

The OOV word rejection capabilities were tested with a rejection database consisting of meeting speech, radio data and car noise, 10 hours each. In all experiments, the system was tuned so that no false alarms were observed with the rejection database. All workstation simulations were in SI domain and the real-time experiment was in SD domain.

### 6.1 Baseline Experiment

First, we carried out a baseline experiment with single-level and whole-phrase HMM approaches. This scheme resulted in average failure rate of over 10% and less than 90% acceptance rate of correct activation phrases. The results can be seen in Table 1.

**Table 1.** Experimental results with single-level recognizer and whole-phrase model.

Environment	One of the three trials	Failures
Clean	79.1%	20.9%
10 dB	88.9%	11.1%
0 dB	96.6%	3.4%
-10 dB	93.1%	6.9%
Average	<b>89.4%</b>	<b>10.6%</b>

### 6.2 Single-Level vs. Multi-Level Approach

In Chapter 3, the motivation for the multi-level recognition approach was explained. In this section, we compare the proposed multi-level approach to the single-level approach. Notice that since each of the test recordings had the activation phrase uttered three times, the single-level acceptance values do not differentiate whether the recognition was accepted with the 1<sup>st</sup>, 2<sup>nd</sup> or 3<sup>rd</sup> trial. For the multi-level experiments, the accepted results for the 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> trial are given separately.

**Table 2.** Experimental results with single-level and subword HMM approach.

Environment	One of the three trials	Failures
Clean	93.7%	6.3%
10 dB	93.5%	6.5%
0 dB	96.7%	3.3%
-10 dB	93.4%	6.6%
Average	<b>94.3%</b>	<b>5.7%</b>

**Table 3.** Experimental results with the proposed approach (multi-level and subword HMM).

Environment	One of the three trials	1 <sup>st</sup> trial	2 <sup>nd</sup> trial	3 <sup>rd</sup> trial	Failures
Clean	99.9%	83.7%	16.2%	0.1%	0.1%
10 dB	100.0%	88.4%	11.5%	0.1%	0.0%
0 dB	100.0%	92.7%	7.2%	0.1%	0.0%
-10 dB	97.4%	87.4%	8.3%	1.7%	2.6%
Average	<b>99.3%</b>	<b>88.1%</b>	<b>10.8%</b>	<b>0.5%</b>	<b>0.7%</b>

Tables 2 and 3 show that the performance of the proposed approach is significantly higher than the performance of the

single-level approach. The proposed method provided 88% error-rate-reduction of the correct activation phrases.

### 6.3 Whole-Phrase vs. Subword HMM Approach

The usage of multiple parts for each activation phrase increases the OOV word rejection capabilities of the system. Three parts per model already gives roughly 27 times better rejection. To verify this, the same experiment was carried out for the whole-phrase HMM approach. Table 4 gives the results of the experiment with the whole-phrase HMM approach.

**Table 4.** Experimental results with whole-phrase HMM and multi-level approaches.

Environment	One of the three trials	1 <sup>st</sup> trial	2 <sup>nd</sup> trial	3 <sup>rd</sup> trial	Failures
Clean	98.8%	71.2%	27.6%	0.0%	1.2%
10 dB	99.7%	79.7%	20.0%	0.0%	0.3%
0 dB	99.6%	92.1%	7.5%	0.0%	0.4%
-10 dB	96.2%	88.4%	7.8%	0.0%	3.8%
Average	<b>98.6%</b>	<b>82.9%</b>	<b>15.7%</b>	<b>0.0%</b>	<b>1.4%</b>

With the whole-phrase HMM approach the average error rate increased from 0.7% to 1.4 % as compared to the proposed approach (see Table 3).

Based on the experiments, we can conclude that both multi-level and subword HMM approaches improve the baseline performance. Combination of the two methods provides the best performance.

### 6.4 Real-Time Experiments

Finally, real-time SD experiments were carried out to evaluate the performance of the proposed approach with novice users. The speakers were from eight different countries and in each language there was a different set of activation phrases. The task of each test person was to utter an activation phrase and if the recognizer did not respond, the activation phrase was repeated. If the recognizer did not respond with three activation phrases, the recognition was discarded and marked as a failure. Otherwise, the recognition was marked as successful. It was also marked, whether the successful utterance was the 1<sup>st</sup>, 2<sup>nd</sup> or 3<sup>rd</sup>. The experiments were carried out in a car. The environments were parked car, city traffic, road 80 km/h, road 80 km/h with music and highway 120 km/h. The results for each environment are shown in Table 5.

**Table 5.** Results of the real-time SD experiments with 12 test persons.

Environment	One of the three trials	1st trial	2nd trial	3rd trial	Failures
park	99.7%	95.8%	3.2%	0.8%	0.3%
city	99.8%	92.1%	6.7%	1.0%	0.2%
road	99.9%	95.1%	4.6%	0.2%	0.1%
road+music	99.1%	86.3%	10.0%	2.8%	0.9%
highway	99.7%	90.3%	8.4%	1.0%	0.3%
average	<b>99.7%</b>	<b>91.7%</b>	<b>6.8%</b>	<b>1.1%</b>	<b>0.3%</b>

The overall acceptance with maximum of three trials was 99.7%. The OOV word rejection capabilities were tested off-line with the same rejection database as used in the previous

experiments. One false alarm per 25 hours of rejection test data was observed.

## 7. CONCLUSIONS

In this paper, we present a novel approach to implement a truly hands-free voice activation system. To meet the high performance requirements for rejecting OOV words and accepting valid activation phrases, we propose new methods to be included in a voice activation system. Subword modeling is shown to enhance the baseline system. These subword HMMs can be constructed from whole-word HMM characterizing the activation phrase. Furthermore, we show that a multi-level approach utilizing repeated unrecognized commands is capable of further increasing the performance of a voice activation system. Our real-time tests and off-line workstation simulations justify the viability of the proposed voice activation approach.

## 8. REFERENCES

- [1] R. Rose, D. Paul, "A Hidden Markov Model Based Keyword Recognition System", *Proc. ICASSP*, pp. 129-132, Albuquerque, USA, 1990.
- [2] J. Wilpon, L. Rabiner, C.-H. Lee, E. Goldman, "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", *IEEE Trans. On ASSP*, Vol. 38, No. 11, pp. 1870-1878, 1990.
- [3] H. Bourlard, B. D'hoore, J.-M. Boite, "Optimizing Recognition and Rejection Performance in Wordspotting systems", *Proc. ICASSP*, pp. 373-376, Adelaide, Australia, 1994.
- [4] J. Junkawitsch, H. Höge, "Keyword Verification Considering the Correlation of Succeeding Feature Vectors", *Proc. ICASSP*, pp. 221-224, Seattle, USA, 1998.
- [5] O. Viikki, K. Laurila, P. Haavisto, "A Confidence Measure for Detecting Recognition Errors in Isolated Word Recognition", *Proc. SST*, pp. 67-72, Adelaide, Australia, 1996.
- [6] K. Laurila, "Noise Robust Speech Recognition with State Duration Constraints", *Proc. ICASSP*, pp. 871-874, Munich, Germany, 1997.
- [7] O. Viikki, D. Bye, K. Laurila, "A Recursive Feature Vector Normalization Approach for Robust Speech Recognition in Noise", *Proc. ICASSP*, pp. 733-736, Seattle, USA, 1998.