

# **NOISE ROBUST VOICE ACTIVATED DIALLING**

**Kari Laurila, Markku Mettälä**

Speech and Audio Systems Laboratory  
Nokia Research Center  
Sinitaival 6, 33721 Tampere, Finland  
Tel: + 358 3 272 58 76  
kari.laurila@research.nokia.com

## **1. INTRODUCTION**

Telecommunications is one of the key application areas for automatic speech recognition (ASR) algorithms. In voice activated dialling (VAD), ASR brings two types of advantages over the conventional modalities. First, ASR improves the convenience of making calls. For example, one has to simply say a name instead of making a sequence of key presses. Second, the usage of cellular phones in vehicles during driving is always a safety risk. Only ASR is capable of providing fully hands-free and eyes-free dialling, enabling drivers to focus more on driving.

Currently, ASR is not yet widely utilized in the call initiation. To a great extent, we still have to blame technology. If we consider car environment (usage of a hands-free microphone, 0 dB signal-to-noise ratio), even a simple name recognition task remains unsolved. So, there is room for improvement, and not least in noise robustness sense.

In the literature, several noise robust feature extraction techniques have been proposed [1-3]. Noise compensation can also be carried out in the recognition unit as in the technique called Parallel Model Combination (PMC) [4], where the HMMs estimated in clean environment are modified to characterize the current noise conditions.

We conventionally achieve noise robustness by means of duration modelling [5] and feature vector normalization [6] schemes. In this paper, we propose a novel user-interface related approach that further enhances noise robustness of VAD. We also propose a true VAD scheme with a simple dialogue structure that enables true hands-free and eyes-free dialling.

## **2. CONVENTIONAL METHODS TO IMPROVE NOISE ROBUSTNESS**

Noise robustness is one of the basic requirements for practical ASR systems. Typically, high recognition accuracy can be obtained if training and testing environments are noise free. In VAD, noise free training and testing environments cannot be generally guaranteed. For example, a hands-free car environment VAD system may have to cope with signal-to-noise ratios (SNR) below 0 dB. It is widely known that mismatches between training and testing conditions, that is, channel variations, different microphones, different background noises and other variabilities present in VAD applications, cause significant reduction in recognition accuracy.

In this chapter, we briefly summarize our noise robust speech recognition approach, which consists of two main parts: duration modelling and feature vector normalization.

## 2.1 Duration Modelling

It is known that standard HMMs are not able to model the temporal structures of speech effectively. Due to this, very different state-frame alignments can be obtained in different noise conditions. In order to minimize deviations between training and testing, we set clear restrictions for possible alignments. In [5], we propose a state duration constrained maximum likelihood (SDML) training, in which we apply duration constraints already in the training phase. In the recognition phase, we use a modified Viterbi algorithm which is performed on a three-dimensional (time, state, duration) space [7]. This way we can always ensure good match between training and testing alignments.

An example state duration constrained HMM structure is given in Figure 1. The filled states represent the conventional left-to-right HMM states and the unfilled states share the same parameters (Gaussian densities) with the filled states in the same vertical lines. One vertical line corresponds to one actual state. State transitions define the allowed state durations. In the example, the first state has the minimum duration of 3 and the maximum duration of 5.

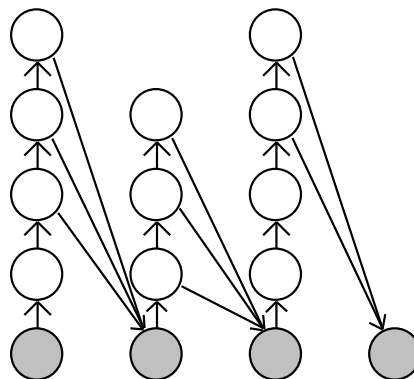


Figure 1: An example state duration constrained HMM structure.

In [5], we showed the effectiveness of the proposed state duration constrained HMMs in a speaker-dependent isolated-word recognition task (e.g. name dialling is typically such). The vocabulary consisted of 30 confusable Finnish first names. Names were spoken by 4 male speakers. In the experiments we used 1-mixture variable-state HMMs for the spoken names and training of each name was done with a single utterance in clean environment. In order to get noisy test utterances, we added car noise to clean utterances with different SNRs\*.

As Table 1 shows, we were able to reduce the error rates in all noise conditions, including clean environment. The most significant error rate reductions, over 80%, were obtained in very noisy conditions having  $-5$  dB or  $-10$  dB SNRs.

\*Throughout this paper SNR is calculated so that DC levels are first subtracted from speech and noise signals and the obtained signal powers are then divided. No spectral shaping is done prior to SNR calculation.

	Clean env.	SNR = +5 dB	SNR = 0 dB	SNR = -5 dB	SNR = -10 dB
Unconstr.	96.25%	92.87%	75.96%	29.77%	8.23%
Constr.	96.59%	94.01%	93.46%	93.47%	86.36%
<b>E.r.r.*</b>	<b>9.07%</b>	<b>15.99%</b>	<b>72.80%</b>	<b>90.70%</b>	<b>85.14%</b>

\*E.r.r. = error rate reduction

Table 1: Duration unconstrained vs. duration constrained HMMs.

## 2.2 Feature Vector Normalization

The other building block in our noise robust speech recognition engine is the so called feature vector normalization. In the current state-of-the-art speech recognition systems, the Mel-Frequency Cepstral Coefficients (MFCC) are widely used to characterize the speech input. Since statistics of the MFCCs vary significantly depending on noise conditions, we proposed in [6] a normalization technique which converts the output of the feature extraction unit to have equal segmental statistics in all noise conditions in order to reduce the mismatch between the training and testing conditions.

The fundamental idea behind the segmental feature vector normalization technique is that irrespective of noise conditions, the output of the feature extraction unit is forced to the same numerical range. The MFCCs are normalized to zero mean and unit variance within a segment of interest.

We carried out a speaker-dependent isolated-word recognition experiment in which we utilized state duration constrained HMMs. Training and testing configurations were the same as in the duration modelling experiments except that the speakers and the vocabularies were different, which explains why there is no exact match or continuation between the result tables.

In the experiment, we compared the performance obtained with unnormalized and normalized MFCCs. By means of normalization we were able to reduce the error rates in all noise conditions. The noisier the environment, the more reduction we obtained. The most significant error rate reduction of 70% was obtained in -10 dB SNR.

	Clean env.	SNR = +5 dB	SNR = 0 dB	SNR = -5 dB	SNR = -10 dB
Unnorm.	96.9%	95.3%	94.0%	89.3%	70.1%
Norm.	97.5%	96.3%	96.1%	94.6%	91.2%
<b>E.r.r.</b>	<b>19.35%</b>	<b>21.28%</b>	<b>35.00%</b>	<b>49.53%</b>	<b>70.57%</b>

Table 2: Unnormalized vs. normalized feature vectors.

### 3. PROPOSED APPROACH FOR VOICE ACTIVATED DIALLING

Like described in the previous chapter, we have been able to gain significant improvements in the recognition rates of a VAD system by enhancing the underlying ASR algorithms. However, we have still not reached the performance we are looking for (who has?). We know that there are many potentially significant algorithm domain improvements that could be done, but in this chapter we want to focus on a user-interface related approach.

The most straightforward improvement in the recognition rates that can be achieved by a modification of a user-interface (UI) is to have multi-utterance training instead of single-utterance training. Based on our recognition experiments, we know that with multi-utterance training we are able to further reduce the error rates by some 40%, but we do not consider this as a very attractive approach.

By viewing the UI issue from another angle, one can say that there are two basic approaches in VAD. That is, implementation with or without a button. We refer to the buttonless approach as a true VAD. True VAD has extremely high requirements since false alarms are considered as very critical errors, to be avoided at any cost. On the other hand, correct activation words spoken by the user should be well detected. It is not easy to meet both of these criterias simultaneously, improving the false alarm aspect degrades the command word acceptance aspect and vice versa. Due to this problematic tradeoff and the high cost of malfunction, true VAD remains as a very challenging task and it is not a wonder that true VAD systems have not been commercially available to the end customers of telecommunications devices so far.

In this paper, we propose an approach which combines button based VAD and true VAD approaches. This means that the user perceives no difference between these two approaches from the ASR point of view. With or without a button the speech dialogue remains the same.

Our approach is targeted to obtain the following:

*Training:* Single utterance training, rejection of invalid utterances, more accurate end-point-detection and enhanced noise robustness.

*Recognition:* Flat dialogue structure (combining voice activation and name recognition), high rejection performance of out-of-vocabulary words, high recognition accuracy.

In the proposed approach an utterance is divided into three sub-units, called *head*, *name tag* and *tail*. Head and tail are always the same and they are used as “markers”, so only the name tag contains unique information.

#### 3.1 Training

While training a new name to the system, the user is required to speak the following utterance:

**HEAD - NAME TAG - TAIL**, e.g. “Call **John Smith** please”

Head and tail parts are localized by means of ASR. That is, an HMM-based speech recognizer finds the most likely positions for the head and tail words as a side product of the recognition and hence end-point-detection (EPD) of the actual name tag is carried out implicitly. EPD by ASR enables us to get rid of the conventional instantaneous signal power and zero-crossings based EPD approach, which is very vulnerable in noisy conditions. In addition, the usage of ASR provides good rejection of invalid training utterances. If the user accidentally says something like “*Oh, what was the name?*”, then the confidences for head and tail parts remain low, which results in a correctly rejected training session.

### 3.2 Recognition

In the recognition phase, the user should follow the same speech dialogue structure as in the training phase regardless of button based or buttonless approach, which is logical to the user. In a true VAD case, the usage of a relatively long and phonetically rich sentence enables us to have good rejection of out-of-vocabulary words. Equal rejection performance could not be achieved with a name tag only. In addition, the usage of head and tail parts enables more accurate localization of the name tag in very noisy conditions.

The proposed approach provides a very simple dialogue structure, combining activation and name recognition phases. Thus, also the interaction time with the system is minimized.

## 4. EXPERIMENTS

One of the main purposes of head and tail models is to improve the localization of the name tag in the recognition phase. In the conventional name recognition approach HMMs are aligned with the input feature vectors and the optimum location (in time domain) of the best match HMM for the name has no restrictions. It is not guaranteed that the name is actually recognized from the spoken name region. It may happen that the best match model is recognized from the background noise and the spoken name is considered as noise (or garbage) by the recognizer.

In order to see how often the best match name model is recognized from the correct location we organized a test. For the test we used an isolated utterance database recorded in clean environment (Finnish first and full names). First we found the end-points (labels) for all spoken names by using automatic EPD algorithm. Then we trained name tags for each speaker using single-utterance training (duration modelling and normalization schemes were utilized). After this we added car noise to the test set utterances with different SNR values. By recognizing the test set utterances we obtained another set of labels produced by the recognizer. Then we compared these label sets and indeed observed some differences.

Figure 2 depicts the average absolute difference of the recognized and end-pointed labels normalized by the duration of each spoken name (duration is obtained from the EPD labels). The position 25% at the x-axis means that labels differ by 25% of the duration of the spoken name. Y-axis gives the relative occurrence in the whole test material.

From Figure 2 one can observe that in clean environment about 5% of the recognized names have 30% difference to the EPD labels. In  $-5$  dB SNR as much as 25% of the recognized names have

30% difference to the EPD labels. One has to conclude that significant amount of recognitions in the conventional isolated word case occur from wrong time domain positions.

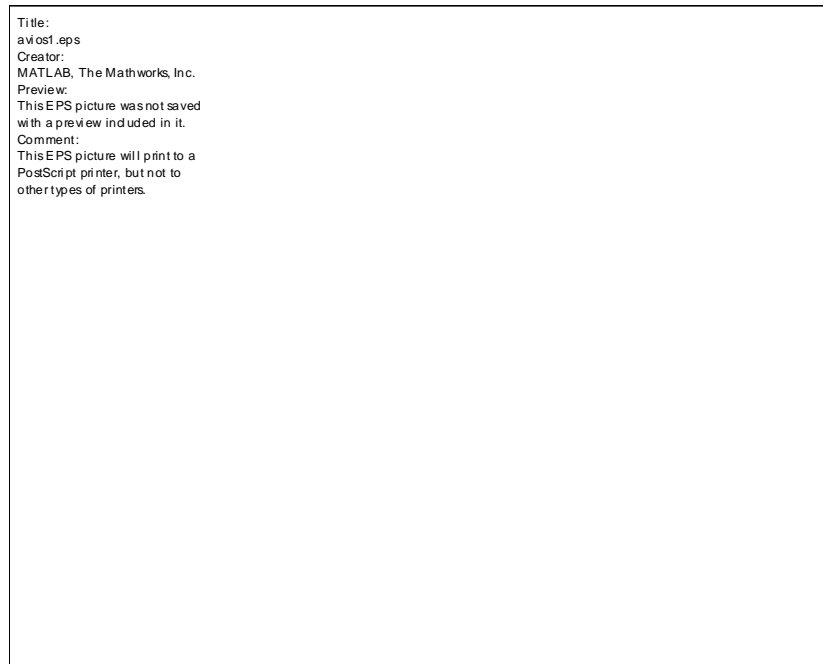


Figure 2: Deviations of the recognized utterance positions as compared to EPD labels (normalized by durations of spoken names).

## 4.1 Training

Training of head and tail models was carried out with isolated utterances. Each speaker uttered head word ("*Soita*") and tail word ("*Puhelu*") once. Afterwards, state duration constrained speaker-dependent HMMs were created out of these utterances.

Name models were trained using the scheme proposed in chapter 2. That is, head and tail models were used to localize each spoken name, and a state duration constrained model was trained from the recognized name region.

Conventional training, in which each name is spoken in isolation and end-point-detected based on instantaneous signal power and zero-crossing values, is rather vulnerable to different background noises, background speech, breath noise and other disturbances. We have observed that SNR should be reasonably high, about 15-20 dB, to ensure that conventional training gives reliable estimates for starting and end points of spoken names. In addition, the conventional method is unable to reject such utterances as "*Oh no!*", that is, a phrase which is something else than an asked name.

Due to the mentioned weaknesses of the conventional training we made a simple experiment to test the robustness of the proposed training method. In the experiment valid and invalid training utterances were given as input to the proposed training module. Invalid utterances consisted of

isolated Finnish names, first and full names (head and tail parts were missing). For a training utterance to be accepted we required that confidences of head and tail models were above a threshold. Confidences were estimated as proposed in [8].

Training was carried out in clean environment and with additional car noise with SNRs of 10 dB, 5 dB and 0 dB. Table 3 shows that the proposed system was able to accept correct utterances with 99.87% rate while rejecting 72.66% of the invalid ones in clean environment. In noisy environments correct utterance acceptance rates decreased slightly while the invalid utterance rejection rates increased.

	Clean env.	SNR = +10 dB	SNR = +5 dB	SNR = 0 dB
Acceptance rate of correct utterances	99.87%	99.50%	98.17%	93.05%
Rejection rate of incorrect utterances	72.66%	83.71%	86.32%	89.21%

Table 3: Robustness of the proposed training method.

## 4.2 Recognition

The proposed VAD approach was compared with the conventional isolated word recognition scheme. In model training, instantaneous signal power and zero-crossing values based EPD method was used in the conventional case whereas in the proposed VAD approach the end-points were localized by means of a speech recognizer, as described earlier.

Training of the models was done in clean environment using only one training sample per model. Two separate vocabularies were used. One with 30 Finnish first names, and another with 30 Finnish full names consisting of first and last names. The database included both isolated names and "*head – name tag – tail*" sentences (same names).

Two different front-ends were used in the experiments. First, we tested standard MFCCs and then we applied feature vector normalization scheme described earlier in this paper.

One-mixture state duration constrained HMMs were estimated for the spoken names. Since each model was trained with only a single utterance, a global variance vector was estimated from all 30 training utterances, separately for each model set. This vector was assigned to each state of each model within the set. In the case of feature vector normalization, a unity variance vector was used.

Test database included speech from 10 persons, 5 female and 5 male speakers, each having 4 recording sessions (A, B, C and D). Training of the models was carried out with the session A data. Each model set was tested with the session B, C and D data. Noise recorded in a moving car was added to the session B, C and D utterances with different SNRs in order to test the performance in noisy conditions.

Table 4 gives the recognition rates in the case of unnormalized feature vectors when only first names were spoken. The proposed method provided lower recognition accuracy in clean environment but improved accuracy in noisy conditions. The average absolute recognition rate improvement over all environments due to the proposed method was 5.4%. When full names instead of only the first names were spoken, the proposed method provided significantly improved recognition accuracy in all environments. As can be seen in Table 5, as high error rate reductions as 60-70% were achieved.

<i>Unnormalized, first names</i>	Clean env.	SNR = 0 dB	SNR = -5 dB
Isolated word recognition	98.19%	83.77%	56.24%
Proposed VAD scheme	97.52%	89.60%	67.38%
<b>E.r.r.</b>	<b>-37.02%</b>	<b>35.92%</b>	<b>25.46%</b>

Table 4: The proposed VAD vs. conventional method, unnormalized features and the first names.

<i>Unnormalized, full names</i>	Clean env.	SNR = 0 dB	SNR = -5 dB
Isolated word recognition	99.67%	94.72%	76.73%
Proposed VAD scheme	99.92%	98.08%	81.03%
<b>E.r.r.</b>	<b>75.76%</b>	<b>63.64%</b>	<b>18.48%</b>

Table 5: The proposed VAD vs. conventional method, unnormalized features and full names.

When feature vector normalization was applied, all the recognition rates improved significantly as it can be seen in Tables 6 and 7. As in the unnormalized feature vector and the first name case, the proposed method provided lower recognition accuracy in clean, but higher when noise level increased. Over all environments, the proposed method provided a marginal absolute recognition rate improvement of 0.81%. With full names the proposed method provided significant improvements in all environments. The highest error rate reductions of 83% were obtained in noisy conditions. The most important observation from Table 7 is that the proposed method provided more than 99% recognition accuracy in all noise conditions.

<i>Normalized, first names</i>	Clean env.	SNR = 0 dB	SNR = -5 dB
Isolated word recognition	98.67%	95.92%	88.18%
Proposed VAD scheme	97.62%	95.70%	91.88%
<b>E.r.r.</b>	<b>-78.95%</b>	<b>-5.39%</b>	<b>31.30%</b>

Table 6: The proposed VAD vs. conventional method, normalized features and the first names.

<i>Normalized, full names</i>	Clean env.	SNR = 0 dB	SNR = -5 dB
-------------------------------	------------	------------	-------------

Isolated word recognition	99.75%	98.24%	95.56%
Proposed VAD scheme	99.83%	99.71%	99.26%
<b>E.r.r.</b>	<b>32.00%</b>	<b>83.52%</b>	<b>83.33%</b>

Table 7: The proposed VAD vs. conventional method, normalized features and full names.

It is rather difficult to interpret the results. First of all, we thought that the proposed method would improve the recognition accuracy significantly in the case of the first name recognition, and especially in noisy conditions. We thought that since the full names are phonetically rich already, head and tail parts may not be needed at all to localize them more accurately in the recognition phase. However, the most significant improvements were obtained with full names, and in the unnormalized feature vector case even in clean environment. The results seem to indicate that our initial thinking was completely wrong. But that has no sense.

After thinking a logical explanation for the unexpected results we ended up with the most obvious reason after eliminating others one by one. In the proposed approach, training of the head and tail models is done with isolated words, but the recognition is done with connected words. Due to this we have a coarticulation mismatch. This mismatch causes especially dynamic feature vector components to differ between training and recognition phases in the word boundaries. Word boundary differences are further emphasized by feature vector normalization (as can be seen by comparing Tables 4 and 6), in which about one second buffer around the current feature vector is utilized in determining the normalization coefficients. For example, in the training phase the head model is followed by silence whereas in the recognition phase the head model is followed by a name tag. This mismatch causes the end frames of the head model to be normalized differently. With this reasoning we are also able to explain why the proposed method improved the recognition accuracy more in the case of full names than in the case of the first names. With full names mismatches in the word boundaries have less effect since the boundaries cover relatively small part of the whole word whereas with short names the word boundaries cover larger part of the whole word. The above reasoning explains also relatively poor rejection of incorrect training utterances in Table 3, since the same mismatch was present in that experiment. To verify the explanation one should train the head and tail models from connected word strings (speaker-dependent or speaker-independent models). We will test this approach in the future.

## 5. TRUE VOICE ACTIVATED DIALLING

One of the biggest problems in automatic speech recognition is the poor rejection of out-of-vocabulary (OOV) words. For a speech recognizer, it is much easier to find the corresponding model if the user speaks a word that belongs to the vocabulary, than to decide if a word spoken by the user belongs to the vocabulary at all.

The poor OOV word rejection causes two different types of problems in VAD. Since the user may not always remember what names have been trained, he/she is likely to utter OOV names every now and then. Or, he/she may say something unintended, for example stutter or sneeze. OOV input often leads to the recognition of the closest match model and thus, possibly to a call to an

unintended party. The poor OOV word rejection also makes it difficult to implement fully hands-free and eyes-free VAD (true VAD). In car environment, the user may drive for hours and make calls every now and then. If the calls are initiated with a true VAD system, the recognizer must be running all the time since there are no push-to-talk buttons or so. For the recognizer true VAD is a huge challenge. It is required that the system should not respond falsely to a general speech, radio program or car noise and still have a high acceptance rate for the correct command words. No false alarms in many hours and 95% acceptance of proper command words can be argued to be a minimum requirement for a true VAD system to be useful.

### 5.1 Isolated Name Recognition as True VAD

The most straightforward idea to implement true VAD is to utilize only the name models. If the user speaks a word that belongs to the vocabulary with high enough confidence (there is a good match between the spoken word and a name model), a call can be made. On the other hand, if the input signal is far away from trained names, the system does not respond. This would be also very user-friendly approach: users should not remember any special command words. Unfortunately, the approach is very vulnerable. First, short names consisting of couple of phonemes or so cannot be distinct enough models that do not provide good match with general speech every now and then. One the other hand, one cannot require users to train very long names only.

To find out the performance of a conventional isolated name recognizer as a true VAD system, we carried out a simple experiment. We included 15 Finnish first names and 15 Finnish full names in the vocabulary. Each speaker had his/her own vocabulary and each name model was trained with a single utterance. The recognizer's ability to accept correct words and reject OOV words was tested with isolated utterances. Each speaker produced OOV utterances that consisted of arbitrary Finnish names that did not belong to their vocabulary.

The acceptance/rejection decision was based on a confidence measure proposed in [8], though it had been enhanced, especially to be SNR-dependent. The test database included speech from eight speakers, recorded in clean environment. Car noise was added to the test set utterances in order to simulate noisy conditions.

Table 8 shows that the conventional name recognizer is unable to meet the requirements of a true VAD system. With over 95% acceptance of correct names in all environments the attained OOV rejection rates of 56%-85% correspond to hundreds of false alarms within an hour of constant speech. Based on this experiment, it is clear that much more enhanced scheme is required in order to meet the OOV word rejection requirements.

	Clean env.	SNR = 0 dB	SNR = -5 dB
OOV word rejection rate	85.15%	56.21%	58.67%
Vocabulary word acceptance rate	98.98%	98.45%	96.10%

Table 8: Isolated name recognition as true VAD.

## 5.2 Proposed Approach for True VAD

Since the confidence of an isolated spoken name is not enough to reach the required OOV word rejection rate, one has to find supplementary schemes. In our true VAD development work we have observed three aspects that are significant from the overall system performance point of view. First of all, activation commands must be phonetically rich in order to differ from general speech. Then, division of overall model into sub-models and utilization of confidences and sequential combinations of the sub-models in the recognition phase greatly enhances the rejection of OOV words. Finally, exploiting repeated spoken unrecognized commands further helps us to increase the OOV word rejection.

One of the most important requirements in true VAD is that the command word must be phonetically rich enough to be distinguished from general speech. Since we cannot require name to be such, one possible solution is to attach a keyword (or keywords) to the name, like in our proposed VAD approach. If the attached keywords are phonetically rich, then the name itself has no high requirements in phonetic sense. It is clear that a combined confidence of many words is better than the confidence of only one of the words.

Division of command word model into sub-models gives us a special opportunity to increase OOV word rejection capability. If  $N$  sub-word models instead of a single model are recognized with a loop grammar, one has  $N^N$  alternative combinations for  $N$  consecutively recognized models instead of only one. If we assume that only one combination is accepted as a command word, then the rejection ability increases approximately by a factor of  $N^N$ , since general speech or background noise is assumed to produce random combinations.

In addition to the above schemes one can still apply a relatively trivial way of improving the rejection of OOV words. One can constantly monitor the sub-word confidences and in "close to acceptance" cases the thresholds can be lowered for a while. In other words, when we think the user may have said a command word that got rejected we lower the thresholds and wait for a repetition (in case of no response, it is natural for a user to repeat the spoken command). Lowering the threshold for some time increases the risk of false alarms but this risk can be eliminated by comparing the assumed successive repetitions. If the repetitions are close to each others according to a certain measure, we can accept the repeated command word. If the repetitions are not close to each others, we may conclude that the first command was not actually a command, but other speech or background noise instead. We call this scheme a multi-level true VAD, where the amount of levels is not limited to two (utilizing one repetition). We are utilizing up to two repetitions, so we have a three-level true VAD system. The overall system can be adjusted so that the desired command word acceptance and OOV word rejection balance is achieved in each level.

In order to test the performance of the proposed true VAD approach we made an experiment with "*head X tail*" utterances in which *head* was "*Soita*", *X* was a Finnish first name and *tail* was "*Puhelu*". Notice that only single utterance speaker-dependent training was used. Notice also that no special head and tail models were trained from isolated utterances, but instead, the proposed true VAD system was trained with complete "*head X tail*" utterances. Due to this, there was no

coarticulation mismatch between training and recognition phases. Training utterances were divided into sub-models regardless of real word boundaries.

The test database included 10 persons, 5 female and 5 male speakers. Each speaker produced 15 utterances in which the "*head X tail*" sentence was repeated three times in order to support the proposed multi-level VAD approach. Again, car noise was added to clean utterances in order to test the performance in noise.

From Table 9 one can see that the first level acceptance rate remains well over 95% in 0dB SNR, and the third level acceptance rate approaches 95% in as low as -10dB SNR.

OOV word rejection performance was tested with speech and car noise input. 10 hours of meeting discussion was given as input to the true VAD system and no false alarms were detected. False alarms were not observed with 10 hours of car noise either.

	Clean env.	SNR = 0 dB	SNR = -10 dB
Single utterance	100%	97.9%	78.9%
One additional repetition	100%	99.8%	91.3%
Two additional repetitions	100%	100.0%	94.2%

Table 9: Command word acceptance rates of the proposed true VAD system.

## 6. CONCLUSIONS

Automatic speech recognition systems are gradually enetering to our daily lives. At this moment however, inadequate recognition accuracy is still one of the most hindering factors. Many ASR systems that are functional in laboratory conditions fail to work in real life environments.

In this paper, we present a voice activated dialling system featuring enhanced noise robustness. We show that by means of duration modelling and feature vector normalization methods we are able to reduce the error rates significantly in noisy conditions. In addition, we propose a voice activated dialling approach which further improves noise robustness and enables true hands-free and eyes-free VAD. By experiments we show that the proposed approach provides more reliable training phase being able to reject incorrect training utterances. We also show that the proposed approach provides improved recognition accuracy while providing a flat dialogue structure for true VAD.

Despite of positive results obtained with the methods presented in this paper, some of the results are rather unexpected and contradictory, revealing that there are still potential ways to significantly improve the proposed VAD approach.

## REFERENCES

- [1] D. van Compernelle, T. Claes, "SNR-Normalisation for Robust Speech Recognition", Proc. of International Conference of Acoustics, Speech, Signal Processing, Atlanta, USA, Vol. 1, pp. 331-334, 1996.
- [2] A. Acero, R. M. Stern, "Cepstral Normalization for Robust Speech Recognition", Proc. of Speech Processing in Adverse Conditions, pp. 89-92, Cannes-Mandelieu, France, 1992.
- [3] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", Journal Acoust. Soc. Am., Vol. 87, No. 4, pp. 1738-1752, 1990.
- [4] M. Gales, S. Young, "Cepstral Parameter Compensation for HMM Recognition", Speech Communication, Vol. 12, No. 3, pp. 231-239, 1993.
- [5] K. Laurila, "Noise robust speech recognition with state duration constraints", Proc. Of International Conference of Acoustics, Speech, Signal Processing, Vol. 2, pp. 871-874, 1997.
- [6] O. Viikki, K. Laurila, "Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization", ESCA-NATO tutorial and research workshop on robust speech recognition for unknown communication channels, Pont-a-Mousson, France, pp. 107-110, 1997.
- [7] H. Gu, C. Tseng, L. Lee, "Isolated-Utterance Speech Recognition Using Hidden Markov Models with Bounded State Durations", IEEE Transactions on Signal Processing, Vol. 39, No. 8, pp. 1743-1752, 1991.
- [8] H. Boulard, B. D'hoore, J.-M. Boite, "Optimizing recognition and rejection performance in wordspotting systems", Proc. of International Conference of Acoustics, Speech, Signal Processing, Vol. 1, pp. 373-376, 1994.