# Detection of voiced segments in noisy speech as step towards a new robust recognition scheme

Andreas Kitzig, Frank Kremer, Hans-Günter Hirsch

*Institute for pattern recognition, Niederrhein University of Applied Sciences, 47805 Krefeld, Germany*
*Email: andreas.kitzig@hs-niederrhein.de*

## Abstract

Nowadays it is possible to achieve good speech recognition rates even in noisy environments. This is made possible through the use of robust speech recognition systems which utilize noise suppression methods or an adaptation of the reference models to the actual noise condition to enhance the recognition rate. But it becomes apparent that there is still a need for improvement when we compare the recognition rates of current robust recognition systems to the accuracy of human speech recognition. Most robust recognition systems show noticeably inferior results compared to the human speech recognition in noisy environments. This is based in a special characteristic of robust human speech recognition. Humans are able to understand the contents of speech in noisy environments by recognizing only certain fragments of the speech and extending this fragmental knowledge for the understanding of the whole utterance. These fragments are characterized by a high signal-to-noise ratio as measure for a high speech level in relation to the level of the background noise. This human characteristic provides a very interesting basis for a new robust recognition approach we are working on. In this paper we want to show some results of our initial investigations to this new approach.

## Introduction

Most robust state of the art speech recognition systems contain additional processing blocks to improve the robustness in bad acoustic conditions. In this task, robustness can be achieved e.g. by using noise suppression methods [1] to extract features that are independent from the acoustic condition or through an adaptation of the reference patterns to the actual noise condition [2]. Compared to the results of a non robust speech recognition system, which only uses the basic processing block for feature extraction and recognition, a noticeable improvement in the recognition rate can be observed, but robust recognition systems still achieve a lower recognition rate compared to the human. We get an explanation of the differing recognition rates by comparing the whole processing sequence of a robust speech recognition system to the human interpersonal communication in a noisy environment.

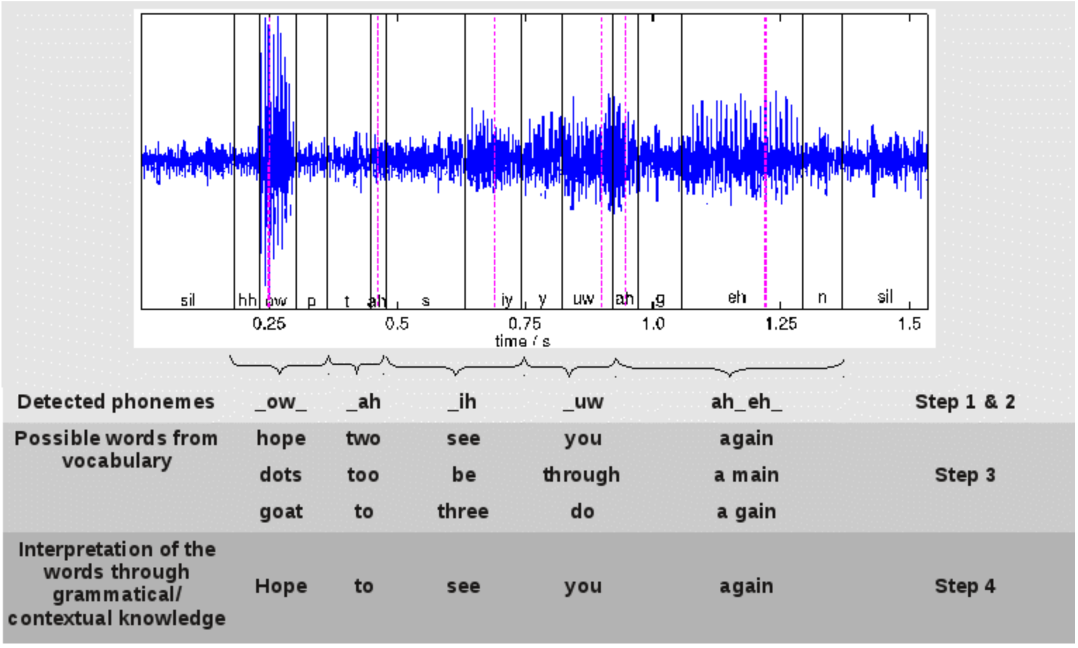
Every speech recognition system starts with a feature extraction. This is done by dividing the whole speech signal into short frames with a length of 20 to 30ms. The frames will be processed in further signal processing steps to obtain relevant acoustic parameters. These parameters are stored in feature vectors for each frame. The resulting temporal sequence of feature vectors is used for the recognition. This is done by calculating the probabilities for the generation of the complete observed sequence of feature vectors through the corresponding set of reference models. In summary, the whole recognition system is working in a straight forward temporal direction. In contrast to the robust recognition systems, humans act different to understand their dialog partner in a noisy environment. A typical scenario for a communication in a noisy environment is shown in figure 1.

**Figure 1:** Example for an interpersonal communication in a noisy environment. The speaker says: ”Hope to see you again!“ while a train is arriving.

Two persons are standing close to a departure platform in a train station and try to communicate while a train is arriving. The speaker says “Hope to see you again” but his or her utterance is disrupted by train noises. So the listener will only get snippets of the conversation, because the parts of the conversation with a low signal to noise ratio (SNR) are not or only partly noticeable to him or her. Speech fragments that contain phonemes with a high speech level are better noticeable to the listener due to a relative high SNR. These phonemes are typically voiced with a periodic signal based on the periodic opening and closing of the glottis.

There are many different approaches to model the human speech recognition process in noisy environments. A simple concept for human speech recognition in noisy environments can be derived from the human characteristic of recognizing only certain fragments of the speech and extending this fragmental knowledge for the understanding of the whole utterance. An overview is given in figure 2.

First, humans try to detect the voiced segments in the noisy utterance. These segments are marked with dashed magenta colored lines in figure 2. In the second step, recognition of the voiced phonemes in the detected segments takes place. At best, the listener understands all voiced phonemes “ow,



| Detected phonemes | _ow_ | _ah | _ih | _uw | ah_eh_ | Step 1 & 2 |
|---|---|---|---|---|---|---|
| Possible words from vocabulary | hope<br>dots<br>goat | two<br>too<br>to | see<br>be<br>three | you<br>through<br>do | again<br>a main<br>a gain | Step 3 |
| Interpretation of the words through grammatical/ contextual knowledge | Hope | to | see | you | again | Step 4 |

**Figure 2:** Steps to describe human speech recognition in noisy environments

ah, ih, uw, ah" and "eh". In step 3, the listener extends the heard fragments to complete words by using his or her vocabulary. For example, the recognize phoneme "ow" is extended to the words "hope", "dots" or "goat". In step 4, the listener is finally able to compose the single words to the whole communicated utterance by using additional grammatical and especially contextual knowledge.

In contrast to the robust speech recognition system that is working in a straight temporal forward direction, this example shows that humans are able to act in several steps to understand the whole content of a dialog in a noisy environment. As mentioned before, it is possible to divide human speech recognition into four rough steps:

1. Detecting parts with high SNR in the utterance
2. Recognize voiced phonemes from the detected parts
3. Extend the phonemes to whole words by using a vocabulary
4. Combine the words to sentences by using grammatical and contextual knowledge

We took these four steps of our human speech recognition concept to derive a new robust speech recognition approach. The new approach resides in the initial phase so we are glad to present the first results from our research. In this paper we want to present an algorithmic description for the detection of the voiced segments and the related results.

## Detection of voiced segments

The initial step of our processing scheme is the detection of speech segments with a high speech level. These segments of the utterance normally contain voiced phonemes. The human processing can be directly transferred to an initial function for the new approach. The difficulty in this task is to ensure a highly accurate detection of the voiced segments in the first processing step to determine reliable segments with a high speech level in the noisy speech signal. These segments are necessary for the further processing without fault in step two and three otherwise the recognition of phonemes or/and the extension to whole words would show wrong results. We investigated three different detection methods and a combination of all three methods to guarantee a reliable detection. We use a robust feature extraction scheme [1] for the calculation of all values that are required for the further processing. This feature extraction function ensures that the further processing is independent from the acoustic condition. A block diagram is given in figure 3.

The first detection method is based on the usage of the short-term energy $\log E$ which can be taken for the detection of voiced segments. Usually, a high energy value indicates a voiced speech segment. To detect voiced segments, the sequence of energy values $\{\log E(1),...,\log E(N)\}$ is smoothed in temporal direction to suppress outlier. In the next step, the actual detection of voiced segments from the energy values is realized. Therefore we implemented a maximum search method which is described in formula 1.

$$peak(n)=\begin{cases}1 \quad for \begin{bmatrix}\log E_s(n)\geq\{\log E_s(n-1),\log E_s(n+1)\}\,\& \\ \log E_s(n)\geq\{thres\}\,\& \\ \log E_s(n)\geq\{mean(\log E_s(1),..,\log E_s(n+1)\}\end{bmatrix} \\ 0 \quad else\end{cases} \tag{1}$$

$$thres=\begin{cases}0{,}99\cdot thres \quad for \quad thres>thres\_init \\ \log E_s(n) \quad for \quad peak(n)==1\end{cases}$$

The initial energy threshold $thres\_init$ which is used for the maximum search is set to a value of 12.

As a second method, we use a voicing information measure. The voicing information can be obtained from the cepstrum. In the cepstrum, the lower cepstral coefficients represent the vocal tract filter and the peak value within the higher coefficients contains information about the fundamental frequency which can be used as a measure for the voicing.

To determine voicing information for every frame of the speech signal, the adaptive filtered DFT spectrum $S(f)$ of the noisy input signal is transferred to the cepstral domain. In a further step, the maximum value of all cepstral coefficients that represent the fundamental frequency in a range from ca. 70Hz to 320Hz is calculated. The maximum value can be used directly as a measure for the voicing. To rate the voiced frames in the speech signal on the basis of the voicing measure, the processing is equal to the energy measure. At first, the values are smoothed in temporal order and in a further step, the maximum search takes place.

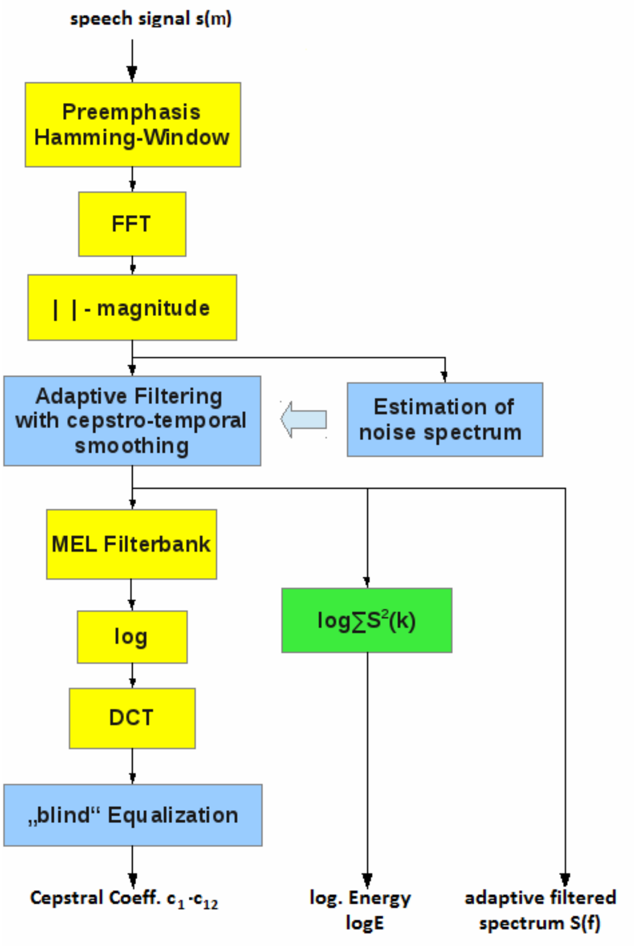

speech signal s(m)
Preemphasis Hamming-Window
FFT
| | - magnitude
Adaptive Filtering with cepstro-temporal smoothing
Estimation of noise spectrum
MEL Filterbank
log
DCT
„blind" Equalization
log∑S²(k)
Cepstral Coeff. c₁-c₁₂
log. Energy logE
adaptive filtered spectrum S(f)

**Figure 3:** Robust feature extraction scheme [1]

The third parameter is a phoneme probability measure. This measure provides a probability value for every frame of the noisy speech signal that indicates whether the frame contains the spectral parameters of voiced phonemes or not. To calculate the phoneme probabilities, a set of phoneme hidden markov models (HMM) is required.

We used an existing set of phoneme HMMs which was trained from the TIMIT [4] database for the calculation of the phoneme probabilities. The HMMs in this set contain three states with 16 mixtures. The phoneme set is aligned to the Carnegie Mellon University (CMU) pronouncing dictionary [3] and contains 39 phonemes. All phonemes in the set are assigned into one of two classes, one class for the voiced and the other class for the unvoiced phonemes. A list of the two classes is given in table 1. The actual calculation of the probabilities takes places after the calculation of a robust feature vector for the current frame.

| | CMU Phonemes |
|---|---|
| Voiced | l, r, w, y, iy, ih, eh, ey, ae, aa, aw, ay, ah, ao, oy, ow, uh, uw, er |
| Unvoiced | m, n, ng, jh, b, d, g, v, z, ch, dh, s, t, zh, f, k, p, sh, th, hh |

**Table 1:** Phoneme Class Assignment

At first the probability for every phoneme HMM in the appropriate class is calculated. We are only interested in the probabilities for the generation of the actual frame through the phoneme HMM in the voiced/unvoiced classes and so there is no need for a temporal modeling. This makes it possible to reduce the complexity of calculation from a HMM to a GMM by using only the middle state with all 16 mixtures of the three state HMM. The resulting probability values for every class $c \in \{voiced, unvoiced\}$ are stored in two vectors $\vec{p}_{voiced}$ and $\vec{p}_{unvoiced}$ and sorted in ascending order:

$$\begin{aligned} &\vec{p}_c = (p_{c1},\dots,p_{cN})\ with\ N = 19, Order \le in\ R \\ &\vec{p}_{c,sort} = (p_{\pi(c1)} \le \dots \le p_{\pi(cN)}), \pi = Perm\ \ of\ \ \vec{p}_c \end{aligned} \quad (2)$$

In a further step, a mean probability value $p_{c,mean}$ for every class is calculated. Therefore the probabilities for the five phonemes of each class that are most probable are used.

$$p_{c,mean} = \frac{1}{M}\sum_{i=1}^{M} p_{c,sort_i} \qquad with\ \ M = 5 \quad (3)$$

To rate the current frame, the two mean probability values are compared to each other. The current frame is labeled as voiced if the mean probability value $\vec{p}_{voiced\ ,mean}$ of the voiced class takes a maximum. Normally, a range of voiced frames is detected. For further processing, only the frame with the highest mean probability value is chosen.

After evaluating the detection rates for the introduced three methods to detect voiced segments in a noisy speech signal it became clear, that the error rates for the single methods are too high for a further processing with regard to a high reliability of the detected segments. Due to this, we combined the detection results of the three methods to force a low error rate. Therefore we choose two simple ways for combination. First, for every section that is detected as voiced by the energy measure, it is checked whether the probability measure and the voicing measure have detected a voiced frame into the same area. In this case, the central voiced frame from the energy measure is taken as voiced. In the case the voicing measure does not show a fitting value, the temporal distance between the center frames of the energy measure and of the probability measure are checked.

If the distance is small, the central voiced frame from the energy measure is taken as voiced frame.

## Results

In this section, we present the evaluation results for the three different methods and the combination in terms of detection and error rates. The detection rate is defined as followed:

$$\text{detection\_rate} = \frac{V - D}{V} \cdot 100\% \qquad (4)$$

$V$ represents the number of voiced segments in the test data. The number of deletions is labeled with $D$ and represents the number of voiced segments that were not detected. The error rate is defined in formula 5. Here $I$ represents the number of segments that were wrongly detected as voiced although they are labeled as unvoiced.

$$\text{error\_rate} = \frac{I}{U} \cdot 100\% \qquad (5)$$

$U$ represents the number of unvoiced segments in the test data. For our experiments we used the training data set of the TIMIT database [4] which contains 4620 speech files with a total number of nearly 57500 voiced and nearly 95300 unvoiced segments. Beside the clean original data we generated four noisy test sets which contain car noises and interior noises in different SNR conditions of 0dB and 5dB. The complete results are shown in figure 4. In general, it can be seen that the detection of the voiced segments works for all three methods almost reliable because all methods achieve a good detection rate. It is obvious that the phoneme probability measure shows the lowest error rate for all three measures. Only the voicing information measure is a bit problematic because it shows a high error rate compared to the other measures. Therefore, the combination of the three measures makes sense. The detection rate for the combination is accordingly lower than the detection rate for the single results but the achieved error rate is close to zero percent. This ensures that a further processing of the detected voiced segments can be done with a high reliability.

## Conclusion and Outlook

As mentioned before, humans are able to understand the contents of speech in noisy environments by recognizing only certain fragments of the speech and extending this fragmental knowledge for the understanding of the whole utterance. In this paper we have presented initial investigations for a new robust speech recognition approach which is oriented at this special human speech recognition concept in noisy environments. As a first step we developed three methods to detect speech segments with a high speech level. The first method obtains a measure from the short-term energy, the second one is a voicing information measure and the third one is a phoneme probability measure that indicates whether the segment contains the spectral parameters of voiced phonemes or not. In several tests we were able to proof that all three methods are working well. To reduce the error rate, we present an additional combination of the methods. In the next step, we are going to implement a phoneme recognizer with respect to step 2 of our human speech recognition concept to recognize the voiced phonemes within the detected voiced segments.

## References

[1] H. G. Hirsch, A. Kitzig, “Robust Speech Recognition by Combining a Robust Feature Extraction with an Adaptation of HMMs”, ITG Fachtagung Sprachkommunikation, Bochum, 2010

[2] Advanced Digital Speech Transmission, Rainer Martin, Ulrich Heute and Christiane Antweiler, John Wiley & Sons, 2008

[3] The CMU Pronouncing Dictionary, v. 0.7a, Reference- URL: `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

[4] TIMIT Acoustic-Phonetic Continuous Speech Corpus, John S. Garofolo, et al., Linguistic Data Consortium, 1993

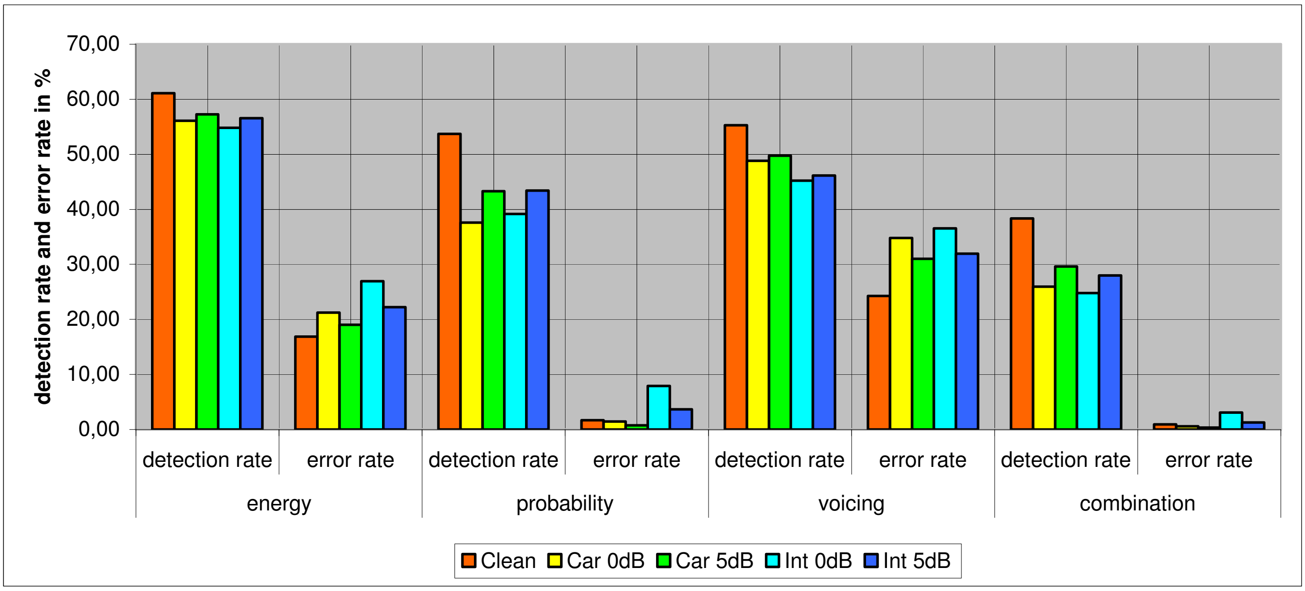


**Figure 4:** Detection results for experiments with different test data