A New Approach to Voice Activity Detection Based on Self-Organizing Maps

Stephan Grashey

Siemens AG, Corporate Technology
Otto-Hahn-Ring 6
D-81730 München, Germany
stephan.grashey@siemens.com

Abstract

Accurate discrimination between speech and non-speech is an essential part in many tasks of speech processing systems. In this paper an approach to the classification part of a Voice Activity Detector (VAD) is presented. Some possible shortcomings of present VAD-systems are described and a classification approach which overcomes these weaknesses is derived. This approach is based on a Self-Organizing Map (SOM), a neural network, which is able to detect clusters within the feature space of its training data. Training of the classifier takes place in two steps: First the SOM has to be trained. When finished, it is used in the second training step to learn the mapping between its classes and the desired output "speech" resp. "non-speech". Experiments on a database containing audio-samples obtained under different noisy conditions show the potential of the proposed algorithm.

1. Introduction

A Voice Activity Detector is an algorithm that deals with the determination of the presence or absence of a voice component in a given speech signal. For example the input of a VAD may be an audio signal recorded from the microphone of a mobile phone. While the user is speaking, the signal is composed of his voice and additional background noise, e.g. traffic noise. The Output of a VAD is a signal, which has additional information, whether the input signal contains speech (i.e. output = 1 or "True") or only noise, i.e. "non-speech" (output = 0 or "False"). Therefore, a VAD can be considered a classifier with $N = 2$ classes.

There are many areas of speech processing where a VAD can be applied: It can be used in data reduction tasks as well as in speech recognition systems, where a better estimation of noise allows for an increase in noise robustness and frame-dropping reduces the computational complexity [1]. Finally in biometric speaker recognition a VAD is needed during enrollment to ensure, that the user's voice and not (more or less big parts) of the background noise is stored as a reference template - otherwise any (other) person having similar background noise during verification would be authenticated.

The problem of deciding between noise only and speech plus noise can be addressed in two different ways: Either by finding optimized features, or by improving the decision strategy of the classifier. Both ways are not independent from each other, of course.

The energy-based approach is a classic one in finding optimized features [2]. It only works well at high signal-to-noise ratios. In early VAD algorithms also zero-crossing rate, autocorrelation function or LPC coefficients were commonly used. In the meantime, spectral entropy [3], cepstral [4], statistical and many other features have been proposed.

Once the decision about the features is made, one can try to optimize the classifier and its decision strategy. Typical approaches are the use of heuristic decision rules, Hidden Markov Models and other statistical classifiers [5], or least-squares classifiers, perceptron, support vector machines and other neural networks [6].

The approach to Voice Activity Detection described in this paper concerns the classification aspects of the two possibilities outlined above: The use of a Self-Organizing Map based on standard features is proposed.

2. Classification approach

2.1. Pre-considerations

As a rule many of the classifiers named above exactly have $N = 2$ classes: One represents "speech", the other "non-speech". Except for the classifiers based on heuristic decision rules, they have to be trained during a (off-line) training phase. Mostly, supervised training is used: The classifiers are trained with "labeled" data, i.e. data which contains additional information whether the current input vector contains "speech" or "non-speech". The labeling has to be done in advance, e.g. manually. Optimization is done by comparing the output of the classifier with the label of the input.

The problem, when taking only two classes from the start is, that a wide range of feature vectors are mapped to the same class, even if their characteristic is totally different. For example, the extracted features for voiced and unvoiced parts of speech (e.g. vocals and fricatives) are probably quite different – but both are mapped to one single class "speech". To overcome these problems, the following considerations are made:

- An appropriate classifier should have much more ($N \gg 2$) than two classes to avoid mapping dissimilar features to only one class. N can be chosen freely, for example $N = 100$. This means, that there are $m > 2$ classes representing "speech" as well as there are $n > 2$ classes representing "non-speech" ($m + n = N$). Thus it becomes possible, to map for example voiced and unvoiced speech segments to different classes.

- The classifier should *decide itself*, whether two feature vectors are similar enough to be put in the same class – or if they are different in such a way, that they have to be mapped to different classes. Of course this idea implies that it is not possible anymore to train the classifier in supervised mode – training has to take place unsupervised.

2.2. Choice of classifier

As described above, we need a classifier, which can discriminate more than two classes, can be trained in unsupervised mode, and should be able to find coherent clusters within the high-dimensional feature-space. There are some potential approaches, which meet these requirements. Here, a Self-Organizing Map (SOM) is used, a widely used approach, which has proven to successfully solve similar problems in other areas (even in speech recognition tasks, see [7] for example).

A Self-Organizing Map is a neural network with one layer of active neurons. The SOM approach was developed by Prof. Teuvo Kohonen in the early 1980s (see for example [8]). It can be viewed as a vector quantization algorithm, which arranges N codebook vectors in a way, that

- the high-dimensional input data space is covered as good as possible.
- a local neighborhood-relationship on a low-dimensional (e.g. two-dimensional) lattice is considered.

Thus a Self-Organizing Map defines a "non-linear projection" of the probability density function of the high-dimensional input data onto a low- (two-) dimensional array. A SOM approximates to its training data during unsupervised training in an ordered fashion and can be used to find relationships with regard to similarity within the high-dimensional feature-space (see Fig. 1).
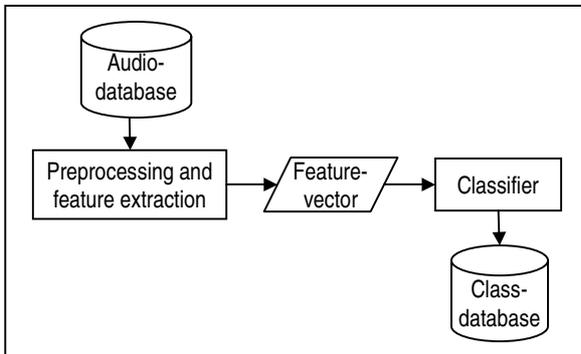

Figure 1: *First training phase: Unsupervised training of the SOM classifier.*

2.3. Mapping of the SOM output

The trained SOM does not contain any information yet regarding the question, whether a particular input vector belongs to "speech" or "non-speech". In the end, of course, this decision is required. Therefore, we need some kind of module, which associates each classifiers class with either "speech" or "non-speech". This module is called "association unit". The "association unit" has to learn its associations during a second training phase. In this second training phase, the SOM runs in classification mode, i.e. as an output to each input vector the corresponding class index is generated. This class index together with the label ("speech" or "non-speech") of the current feature vector forms the input of the "association unit" (see Fig. 2).

The "association unit" now performs a labeling of the class indices: For each class index the number of "hits" of both, the "speech" and the "non-speech"-label is counted. After finishing the training, the label holding the majority is assigned to the particular class index.
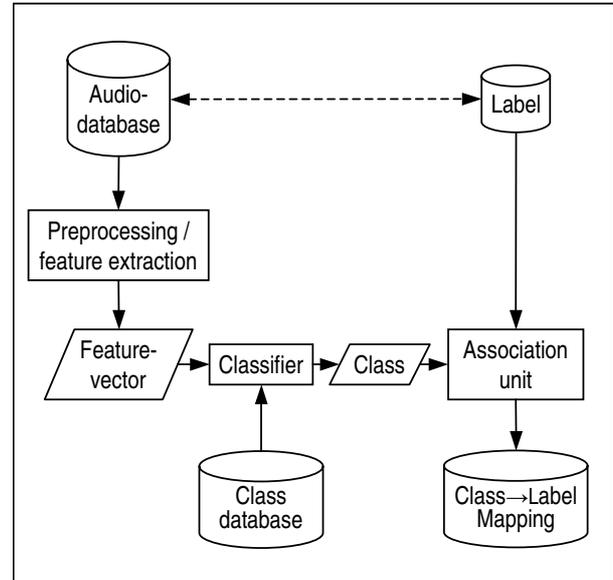

Figure 2: *Second training phase: Training of the association unit (Learning the mapping of each class index to one of the labels "speech" or "non-speech").*

2.4. Using the trained classifier

After successfully finishing both training steps the classifier can be used to determine the presence or absence of speech in an unknown audio signal (Fig. 3): The class index corresponding to each single feature vector is computed by the SOM, and the association unit maps this class index to either "speech" or "non-speech".
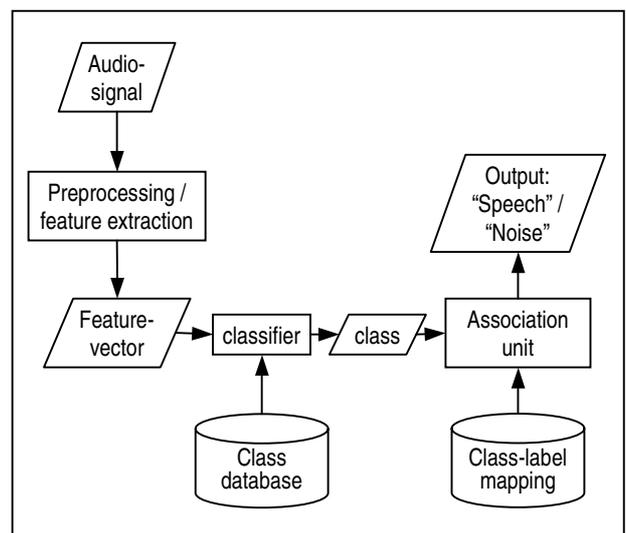

Figure 3: *Usage of the trained SOM-VAD to classify unknown speech samples*

3. Experimental results

3.1. Database

As an evaluation database a subset of the Spanish SpeechDat-Car Aurora Database with a sampling frequency of 8 kHz / 16 bit was used [9]. This database contains 4914 recordings of more than 160 speakers done within a car environment. The recordings include isolated digits, credit card numbers and 6-digit PIN-codes. Like in other SDC-Aurora databases, data was obtained under three noisy conditions and with two different microphone types:

- Quiet (16 %), low noisy (49 %) and high noisy (35 %) conditions, depending on the driving situation.
- Close-talking microphone and hands-free microphone.

3392 files (69 %) were used for training of the classifier, 1522 files (31 %) were used for testing purposes.

3.2. Pre-processing

The input vector of the SOM was formed using 12 mel-frequency cepstrum coefficients (MFCC's): The pre-emphasized speech was framed into lengths of 32 ms with a frame-shift of 15 ms. No context information was used, i.e. each frame was computed independently from its neighboring frames.

3.3. Configuration of the SOM

The SOM used during the experiments had a size of $25 \times 25 = 625$ nodes. A size of 50×50 nodes was tested as well in one experiment, producing slightly better results, but requiring much more training expenditure. A rectangular topology of the SOM was chosen due to its easier implementation than for example a hexagonal or even a dynamic structure. The time needed for the two-stage training with 3392 files resulting in 439590 feature vectors was about 20 minutes on a machine with a Pentium III-Processor / 700 MHz.

3.4. Results

As mentioned above, the SOM and the association unit was trained using 69 % of the files of the database. The remaining 31 % of the files were used to examine the performance of the proposed approach: For this purpose, each decision of the classifier was compared with the label of the corresponding audio frame. These labels were generated before automatically by a Hidden-Markov-Model trained on several databases with fixed telephone data; the labeling was done only on the close-talk data. The results of this comparison are shown in Table 1 and Table 2: In particular a non-match rate of 16 % with frames labeled as "speech" by the HMM but classified as "non-speech" by the SOM can be observed.

To assess these results, it is necessary to have a closer look on the audio signal, on the related reference label of the HMM and on the output of the SOM-VAD: Figure 4 shows as an example the

Table 1: *Comparison of the classifier output with the reference label of the input vector during testing.*

Total number of frames	439590	100 %
Matching frames	350166	80 %
Non-Matching frames	89424	20 %

Table 2: *Detailed view of the non-matching frames from Table 1. Please note the explanations in the text.*

Total number of non-matching frames	89424	20 %
Label: non-speech – classified as speech	17997	4 %
Label: speech – classified as non-speech	71427	16 %

audio signal of a user speaking "cuatro" (Spanish "four"), and, below the signal, the label of the HMM as well as the decision of the SOM-VAD: One can see, that the reference label of the HMM forms a compact block classified as "speech", covering the whole utterance. It can also be seen, that the HMM tends to mark the signal as "speech" even some frames before the actual "speech" starts. On the other hand, the SOM-VAD is very accurate: It classifies as "speech" not until the speech actually starts (perhaps it is a little too late) and it even grasps the small speech pause (= "non-speech") in-between the syllables.
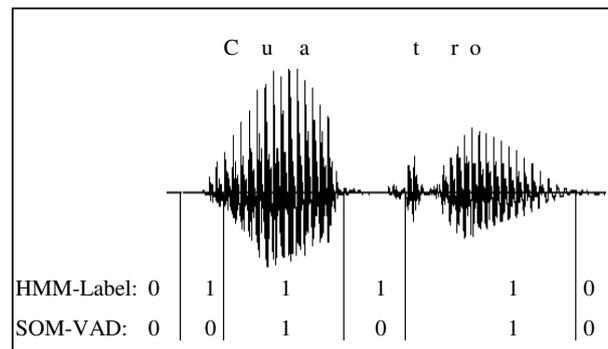

Figure 4: *Example for the Spanish word "cuatro" (four): Comparison of the Labels generated by a HMM-Model and the Output of the SOM-VAD (0="non-speech", 1="speech").*

Now, who made an error? Is the reference label wrong, because it is to liberal when labeling "speech". Or is the decision of the SOM-VAD wrong, because it is to accurate? This shows the conflict, when trying to evaluate one error-prone method with another error-prone method: The error made can not be less than the error already contained in the reference. These aspects has to be considered when looking at the number of non-matching frames.

To adapt to the behavior of the HMM to ignore small pauses within words the output of the SOM-VAD was smoothed by a moving mean filter. The results obtained are shown in Table 3 and Table 4: The number of non-matching frames could be reduced from 20 % to 14 %. Still the major error is made with frames labeled as "speech" by the HMM, but classified as "non-speech" by the SOM-VAD. As the effect of small speech pauses in-between the syllables was reduced to nearly zero by the filtering, these non-matches mainly occur at the transition from "non-speech" to "speech" and from "speech" to "non-speech", i.e. at the beginning and at the ending of an utterance. In these points, it is very difficult to decide, who is making the mistake - the HMM or the SOM-VAD (probably they are both making a mistake). Therefore, the overall results gained by the approach presented in this paper are very promising.

Table 3: *Comparison of the classifier output smoothed by a moving mean filter with the reference label of the input vector.*

Total number of frames	439590	100 %
Matching frames	378365	86 %
Non-Matching frames	61225	14 %

Table 4: *Detailed view of the non-matching frames from Table 3. Please note the explanations in the text*

Total number of non-matching frames	61225	14 %
Label: non-speech – classified as speech	6790	2 %
Label: speech – classified as non-speech	54435	12 %

4. Conclusions

In this paper an approach for the classification module of a Voice Activity Detector is presented. It is based on a Self Organizing Map, a neural network which is "specifically tuned to various input signal patterns or classes of patterns in an orderly fashion" [7].

To avoid the mapping of totally different input features to only one class the SOM contains a great number of nodes and therefore a great number of classes. The training process takes place in two steps. First the SOM is trained in unsupervised mode to find clusters in the high dimensional feature space. In the second step, each class of the SOM is assigned a label representing either "speech" or "non-speech". This is done by training a "association unit". The proposed approach holds some advantages:

- Dissimilar feature vectors are not "forced" into a single class - rather the class is assigned based solely on a similarity criterium. This leads to increased classification accuracy.

- Inexactness during the labeling of the audio data used during the training process has only little effect on the classification result since unsupervised training is performed, and the decision of the association unit is based on a majority criterium. Therefore, even short pauses in between the syllables are correctly classified as "non-speech", though the input typically is labeled as "speech".

- The approach is independent of a language or the content of the input signal.

- Altogether, the accuracy of the VAD is improved, which should lead to a better performance of the following modules depending on the VAD results.

The first results obtained show the potential of the approach. As nearly no optimizations to the SOM has been made so far, there should be still enough potential in improving the results.

5. Acknowledgments

6. References

[1] Astrov, S. and Andrassy, B., "Large Vocabulary Isolated Word Recognition for Embedded Systems", to be presented at Eurospeech 2003, 1-4 September, Geneva, Switzerland.

[2] Ganapathiraju, A., Webster, L., Bush, K. and Kornman, P., "Comparison of Energy-Based Endpoint Detectors for Speech Signal Processing", Proceedings of the IEEE Southeastcon, pp. 500-503, Tampa, Florida, USA, April 1996.

[3] McClellan, S. and Gibson, J.D. "Variable-Rate CELP Based on Subband Flatness," IEEE Trans. on Speech and Audio Processing, vol. 5, pp. 120-130, March 1997.

[4] Haigh, J. and Mason, J., "Robust voice activity detection using cepstral features", IEEE TENCON, pp. 321–324, China, 1993.

[5] Sohn, J., Kim N. and Sung W., "A statistical model-based voice activity detection", IEEE Signal Processing Letters, vol. 6, pp. 1–3, 1999.

[6] Stadermann, J., Stahl, V. and Rose, G., "Voice Activity Detection In Noisy Environments", European Conference on Speech Communication and Technology, Aalborg, Denmark, September 2001.

[7] Honkela, T., "Self-Organizing Maps In Natural Language Processing", PhD thesis, University of Helsinki, Neural Networks Research Centre, 1997.

[8] Kohonen, T., "Self-Organizing Maps", Springer, Berlin, Heidelberg, 1995.

[9] URL: <http://www.speechdat.org/SP-CAR/>