Acoustic Feature Selection for Speech Detection based on Amplitude Modulation Spectrograms

Denny Schmidt, Jörn Anemüller*
Medical Physics Section, Carl von Ossietzky University Oldenburg, 26111 Oldenburg, Germany
{*denny.schmidt, joern.anemueller*}*@uni-oldenburg.de*

Introduction

As acoustic devices possess evermore computing power, signal processing is influenced increasingly by machine learning techniques. E.g., hearing aids detect different listening situations by extracting several features (spectrum, modulations...) and feeding them into a classifier. The question is how to determine features that result in best classification performance and good generalization to new signals.

Here, a feature selection strategy for automatic speech/non-speech classification based on the support vector machine algorithm (SVM) is presented. Input for the selection algorithm is the psychoacoustically motivated amplitude modulation spectrogram (AMS) presented by Kollmeier and Koch [5]. Classification is performed using clean speech signals in the speech class and „clean“ signals from noise-like acoustic scenes in the non-speech class. Results are presented regarding number of modulation frequencies required for speech detection, corresponding classification accuracy and generalization to novel data. Relevant modulation frequencies for speech detection are identified and related to psychophysical evidence.

Amplitude Modulation Extraction

Drullman et al. showed that modulation frequencies f_m in the range of 2Hz up to 8Hz are important for speech intelligibility [2]. Motivated by this and other psychophysical and physiological findings, the chosen front end for the presented feature selection is the AMS which is a representation of modulation intensity in dependence on center frequency f_c and modulation frequency f_m as a function of time. Figure 2 shows the mean AMS of clean English speech reflecting the typical modulation spectral shape of speech with a maximum around $f_m = 3$Hz as described in Houtgast and Steenken [4]. It is expected that modulation frequencies near this maximum are the most salient ones for speech/non-speech classification.

Classification and Feature Selection

Support vector machine classification implemented in the LIBSVM toolbox [1] is used. Features employed here are either different modulation frequency (f_m) bands (exp. 1) or individual (f_c, f_m) bins (exp. 2, 3). Feature selection (cf. Fig. 1) is performed using standard sequential forward selection (SFS) which is based on iterated evaluation of classification performance with different feature subsets. SFS begins with a small (typically empty) feature subset and then adds those features to this subset that maximize classification accuracy in the present iteration. This procedure is iterated and identifies the most salient feature subset for a given number of features. To avoid overfitting, classification accuracy during each stage is evaluated with five-fold cross-validation, which partitions training data into five parts (“folds”), trains five classifiers on the training data, each using a different four-fifth of the data for training and the remaining fifth for accuracy evaluation. Cross-validation accuracy (CV_{ACC}) is then determined as the mean accuracy attained by the five classifiers on the respective test folds (the fifth folds).

Data, Experiments and Results

The training set includes clean speech from the TIMIT database of continuous English speech of different dialect regions. For training the dialect region 1 is used. The non-speech class contains recordings from a street scene recorded at a busy intersection in a distance of 3 meters from the road. The test set is composed of the TIMIT test set of dialect regions 1 and 2 and a recording of a street scene done in close distance to the street. Parts of the NOISEX noise signal database (files „volvo“ and „factory1“) are used as test sets, as well. Sampling rates are 16kHz (exp. 1) and 8kHz (exp. 2 and 3). AMS pattern computation was performed using an STFT decomposition (4ms window length, Hanning window, 3.75ms overlap); a successive bark-band summation and log-compression of amplitudes; and a second stage of STFTs applied to each center frequency channel (1.024s window length, Hanning window, 0.5s overlap). SVM classification was performed with a linear kernel and a hard decision margin.

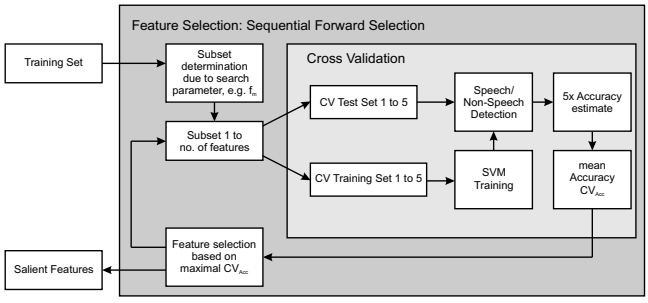

Figure 1: Feature selection scheme for the Sequential Forward Selection algorithm

*This research was supported by the EC under the DIRAC integrated project IST-027787.

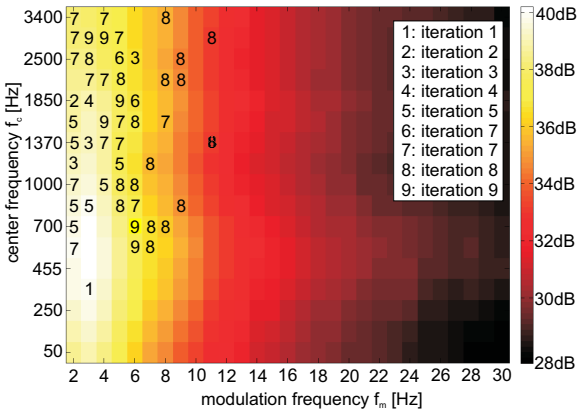

Figure 2: Mean AMS pattern of clean speech training data. Modulation intensity on a logarithmic scale according to the color bar on the right. Numbers denote selection iteration of salient f_m/f_c combinations.

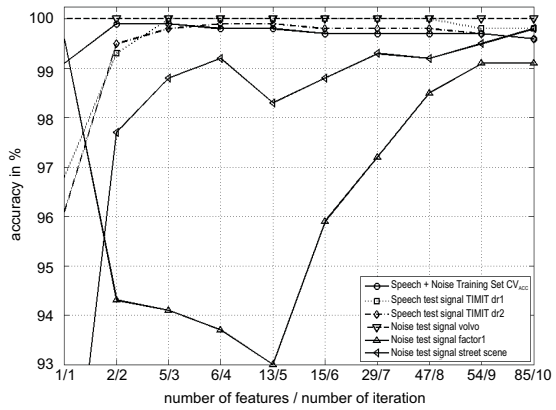

Figure 3: Cross validation accuracy CV_{Acc} and classification accuracy for different test signals for the first 10 iterations of feature selection in percent

Experiment 1. To determine which modulation frequency bands (irrespective of center frequency) are most salient, the entire f_m-band is the search parameter of the feature selection stage. The f_m-band selected as the most salient one (i.e., first selected f_m-band) is the 3Hz-band with a cross-validation classification accuracy of $CV_{Acc} = 99.4\%$. After 9 iterations of including additional f_m-bands, CV_{Acc} reaches its maximum value at 99.8%. This subset contains modulation frequencies $f_m = 3, 4, 26, 25, 9, 14, 28, 13, 20$Hz. Hence, the number of classification features decreases from 493 values of the complete AMS pattern to 153 values in the subset while at the same time accuracy remains essentially unchanged compared to using the entire AMS pattern for classification (which yields $CV_{Acc} = 99.7\%$).

Experiment 2. To reduce the number of features further, the search parameter of the feature selection strategy is the most salient combination of f_c and f_m, i.e. the value of the modulation intensity in single (f_c, f_m)-bins. Fig. 2 displays the optimal features found with 9 iterations of feature selection, each indicated by the iteration number of its selection. The corresponding classification accuracy (cf. "Training Set" curve in Fig. 3) reaches $CV_{Acc} = 99.7\%$ after 9 iterations (54 features). Values of $CV_{Acc} = 99.9\%$ are attained for as little as 2 features Hence, classification accuracy changed only slightly as a result of the large reduction in the number of employed features compared to exp. 1.

Experiment 3. How does the number of selected features influence generalization ability to data from new acoustic environments? To this end, performance of classifiers trained under exp. 2 (and an additional 10th iteration) for different numbers of features/iterations is evaluated on data from previously unheard sound scenes. Results (cf. Fig. 3) show that very low features numbers (below about 6 features) have a tendency to result in poorer classification accuracy on new data. Feature numbers of more than about 50 lead to decent classification accuracy on all test sets. In the intermediate range (about 13 to about 47 features), performance strongly depends on the chosen data set, with generally better performance on speech test data than on noise signal test data. A special case is the exceptionally stationary in-car recording (test signal "volvo") which results in accuracy 100% already for a single features.

Conclusion

A feature selection scheme for speech/non-speech detection based on the amplitude modulation spectrogram was presented. It was shown that feature selection applied to AMS patterns reduces the number of features for speech detection while essentially retaining optimal classification accuracy. The modulation frequency range important for speech intelligibility ($f_m = 2$Hz$\ldots 8$Hz) was determined as being also the most salient for speech/non-speech classification, confirming our expectation. However, modulations in center frequency bands below about 300Hz have failed to show up as particularly salient for speech/non-speech detection. This may be a result of the non-speech class used for training and feature selection (road traffic noise) with its strong low-frequency characteristics. Further, it has been shown that generalization to new acoustic environments may benefit from using more features than cross-validation on training data would suggest.

References

[1] Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines, 2001, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[2] Drullman, R. et al.: Effect of temporal envelope smearing on speech recognition, J. Acoust. Soc. Am., 95(2), 1994

[3] Guyon, I., Elisseeff, A.: An Introduction to Variable and Feature Selection, J. Mach. Learn. Res. 3, 1157-1182, 2003

[4] Houtgast, T., Steeneken, H.J.M.: A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria, J. Acoust. Soc. Am., 77(3), 1985

[5] Kollmeier, B. and Koch, R.: Speech Enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction. J. Acoust. Soc. Am. 95(3), 1994