Detection and Identification of Rare Audiovisual Cues

Insperata accident magis saepe quam quae speres.
(Things you do not expect happen more often than things you do expect) Plautus (ca 200(B.C.)

Project no: 027787

DIRAC

Detection and Identification of Rare Audio-visual Cues

Integrated Project
IST - Priority 2

DELIVERABLE NO: D2.5
Acoustic Classification Method

Date of deliverable: 30.6.2007
Actual submission date: 2.8.2007

Start date of project: 01.01.2006 Duration: 60 months

Organization name of lead contractor for this deliverable: **Carl von Ossietzky University Oldenburg**

Revision [draft, 1, 2, …]

Project co-funded by the European Commission within the Sixth Framework Program (2002-2006)		
Dissemination Level		
PU	Public	
PP	Restricted to other program participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	X
CO	Confidential, only for members of the consortium (including the Commission Services)	

Deliverable D2.5:
Acoustic Classification Method

Carl von Ossietzky University Oldenburg

A classification method is presented that discriminates speech from non-speech sounds and that detects the presence of speech in a background of non-speech sounds. Features used for classification are modulation components of the signal extracted by computation of the amplitude modulation spectrogram. By construction, they are largely invariant with respect to spectral changes in the signal, thereby allowing us to separate the modulation information from purely spectral information. Feature selection techniques and support vector classification are employed to identify the modulation components that are most salient for the classification task and therefore can be considered as highly characteristic for speech.
Results show that highly reliable discrimination and detection of speech can be performed with less than 10 optimally selected modulation features, the most important ones of which are located in the modulation frequency range below 10 Hz. Increasing the number of selected features to about 40 is beneficial for stable generalization to unseen data. Detection of speech in a background of non-speech signals can be performed with more than 90% accuracy for signal-to-noise-ratios (SNRs) down to -10 dB. These results demonstrate the importance of the 2 Hz to 10 Hz modulation frequency range for speech detection and corroborate the significance of modulations in speech pointed out in the literature.

Contents

1 Introduction

Knowledge of the present acoustic environment is a critical piece of information for priming other parts of a multi-level processing chain. E.g., "low-level" audio signal enhancement depends on knowledge of scene characteristics to optimally enhance desired signal components, speech recognition needs to be informed about the presence of speech, and audio-visual recognition algorithms can adapt their priors accordingly. Speech has a special role among acoustic stimuli as it indicates presence of a localized sound source (triggers spatial audio processing) from which information can be extracted (triggers speech recognition) and which forms certain expectations about visual cues (primes visual recognition).

Amplitude modulation features with spectral invariance

One obvious cue to determine present acoustic context is the spectral content of the signal. However, to obtain algorithms robust to realistic variations it has to be taken into account that the recorded spectrum is a function of several parameters, among them source characteristics, environmental acoustic parameters, location of microphones and physical parameters of the recording equipment. Changes in these parameters influence the obtained signals in a way that is hard to compensate for without specific prior knowledge. A multitude of approaches to tackle these problems have been developed, but the variability in realistic environments remains to have a significant impact on, e.g., the performance of speech recognition schemes.
Temporal information represents information that can to a large extent be regarded as independent of spectral content. The analysis of temporal structure of signal envelopes within spectral sub-bands leads in a natural way to the notion of amplitude modulations in individual spectral bands, which in essence characterize periodic or semi-periodic patterns of spectrally confined energy changes. The amplitude modulation spectrogram [7], the computation of which will be described in detail below, represents one way of capturing the modulation structure itself as a function of time, hence the term "spectrogram". It quantifies modulation power computed over a short temporal window (typically one second or less in length) in dependence on spectral position—the "center frequency"—and inverse time-scale of power modulations—the "modulation frequency".
When used with appropriate scaling as detailed below, analyzing temporal structure with the amplitude modulation spectrogram poses the advantage of entirely separating spectral (in the sense of center frequency) from modulation-spectral (in the sense of modulation frequency) content, since the average spectrum (over one observation interval) influences only the values of the modulation spectrogram in the lowest modulation frequency bands. The remaining parts of the modulation spectrogram are invariant to spectral coloring of the sources and spectral shaping introduced by recording equipment. Echoes and reverberation with longer time-constants though will be reflected in the modulation spectrogram, essentially through their effect of reducing modulation depth.

Modulations in speech analysis, psychoacoustics and speech recognition

The advantages of using modulations as robust features might not have gone "unnoticed" by biological systems. Speech as likely our most important communication signal is characterized by modulations that stem from its syllabic structure with semi-periodic signal envelope minima, visible as a peak around $f_m = 3$ Hz modulation frequency [5]. Modulations in the f_m frequency

range from 2 Hz to 8 Hz are of particular importance for (human) speech intelligibility [3]. Automatic speech recognition benefits most from information coded in the modulation frequency range between 1 and 16 Hz, with the dominant component around 4 Hz [6].
The present study focusses on identifying the modulation frequencies that are most relevant for the discrimination of speech from other sounds, thereby aiming to corroborate the previous results on the importance of modulations as a characteristic and significant property of speech. It is expected that modulation frequencies near the maximum around 3 or 4 Hz are the most salient ones for the discrimination between and speech and non-speech and for the detection of speech in realistic ambient acoustic backgrounds.
Modulations and other features have been employed for discrimination of speech and other acoustic classes by several authors. Ostendorf [11] used modulation spectra in three bands to discriminate between three classes, speech, speech in noise and noise. Nordqvist [10] discriminates between three types of acoustic scenes, speech in traffic noise, speech in babble and clean speech, based on cepstral derivatives which can be interpreted as the cepstral transform of modulation band-pass filtered signals, i.e., implicitly a single modulation band is used. Büchler's [1] study for sound classification employed modulations together with several features motivated by auditory scene analysis such as spectral profile, harmonicity and onsets. Mesgarani [9] performs speech discrimination based on auditory model output that employs spectro-temporal receptive fields to capture spectro-temporal modulations.

2 Methods

2.1 Extraction of amplitude modulation spectrogram features

The amplitude modulation spectrogram is extracted in two steps. A first short-term spectral decomposition is applied to yield the standard magnitude spectrogram of the signal. Magnitudes are scaled logarithmically to transform multiplicative factors in each band (corresponding to a convolution of the time-domain signal) to additive terms. The second spectral analysis step extracts modulation spectral information by computing the windowed short-term spectra of the log-scaled signal envelope within each band of the first spectral transform with an analysis window that spans several analysis windows of the first spectral transform.
Standard parameters used in this study are a 32 ms Hanning analysis window for the first spectral analysis with subsequent summation into 17 Bark-scaled bands with center frequencies from 50 Hz to 3400 Hz. The second (modulation) spectral transform employs a 1 s long Hanning window. Hence, within each Bark-(center-)frequency band modulations are determined in 1 Hz steps, and one amplitude modulation spectrogram (AMS) pattern is derived every 500 ms, using a window-overlap of 50% for the second transform.
Invariance with respect to spectral coloring is obtained by excluding the two lowest modulation spectral bands (0 Hz and 1 Hz) from further analysis since any constant additive term on the log-scaled sub-band envelope impacts only those two bands (when using no zero padding for the modulation spectral analysis). Constant here refers to a time-scale of 1 second, leaving invariance also intact for slow drifts in the spectral source characteristics and/or changing physical properties of the overal signal propagation and recording system.
The highest modulation frequency taken into account (after initial experiments) is limited to 30 Hz. The total number of available features is therefore 493: 17 Bark center-frequency bands $\times$ 29 modulation-frequency bands. A full 493-dimensional feature vector is computed every

500 ms. Examples of time-averaged modulation spectrograms are shown in Fig. 2.

2.2 Feature selection

Feature selection has been pursued with the goal to identify which parts of the amplitude modulation spectrogram are most salient for discrimination of speech from other acoustic stimuli and for the detection of speech within realistic acoustic background sounds. In a first approach, entire "vertical slices" from the modulation spectrogram have been considered, each of which holds the modulation information corresponding to a single modulation frequency and all 17 center-frequency bands. Feature selection was used to identify which of these 29 "slices" lead to highest cross-validation accuracy.
In a second, more fine-grained second approach, individual center-frequency/modulation-frequency points are selected from the amplitude modulation spectrogram to maximize classification accuracy. Feature selection in this case can choose from all of the 493 features independently.
Two standard greedy algorithms for feature selection were used, the sequential forward selection (SFS) and sequential floating forward selection (SFFS, [12]) schemes. Starting with an empty subset of used features, SFS sequentially adds those features to the subset that maximize accuracy when added to the already selected subset of used features. If during a forward selection step several features lead to the same performance, the feature to add to the subset of used features is chosen at random from them. SFS monotonically increases the size of the feature subset, but never decreases its size. The procedure can lead to a very suboptimal feature subset since it only searches a very small subspace of feature combinations compared to exhaustive search, and systematically misses all instances where exclusion of a previously selected feature leads to an increase in accuracy in a later iteration.
Sequential floating forward selection adds a backtracking stage to SFS that after each inclusion of a new feature searches "backwards" for features whose exclusion results in an increase in accuracy. Back-tracking is continued as long as exclusion of features increases accuracy, subsequently the algorithm proceeds with the next feature inclusion step. Heuristically, SFFS often leads to better selected feature subsets than SFS. The drawback is its increased computational demand, which is at least twice that of SFS (in the case of all backtracking steps being rejected) and which typically can be about 10 times that of SFS. In our application, SFFS has been found to on average increase accuracy slightly, with an estimated 3-fold increase of computational load compared to SFS, indicating that comparably few feature exclusion steps led to improved classification. Nevertheless, SFFS is clearly a less suboptimal approximation to exhaustive search than SFS and has been the preferred feature selection algorithm for many of the experiments reported here.
Filter methods of features selection, e.g., projection onto a lower-dimensional principal component analysis (PCA) basis or projection based on linear discriminant analysis (LDA) can serve as an alternative means of reducing feature vector dimensionality. This approach was not pursued here since our interest was focused on determining the most salient individual parts of the modulation spectrogram rather than a continuous weighting of all of its parts.

2.3 Classification and performance evaluation

The classification backend employed consists of a standard support-vector-machine classifier [2]. Classifier accuracy has been determined during the feature selection stage using five-fold

cross-validation. Cross-validation folds have been chosen as contiguous parts of the training data. Using randomly sampled folds, as implemented in some toolboxes, resulted in artificially high cross-validation accuracy since consecutive spectra and modulation-spectra of audio data cannot be regarded as independent samples from the data distribution. After training and feature selection, classifier accuracy has also been determined on several test data sets that were taken from a different part of the recordings than the training data, or that originate from different recordings, different scenes or a different data sources than the training material.

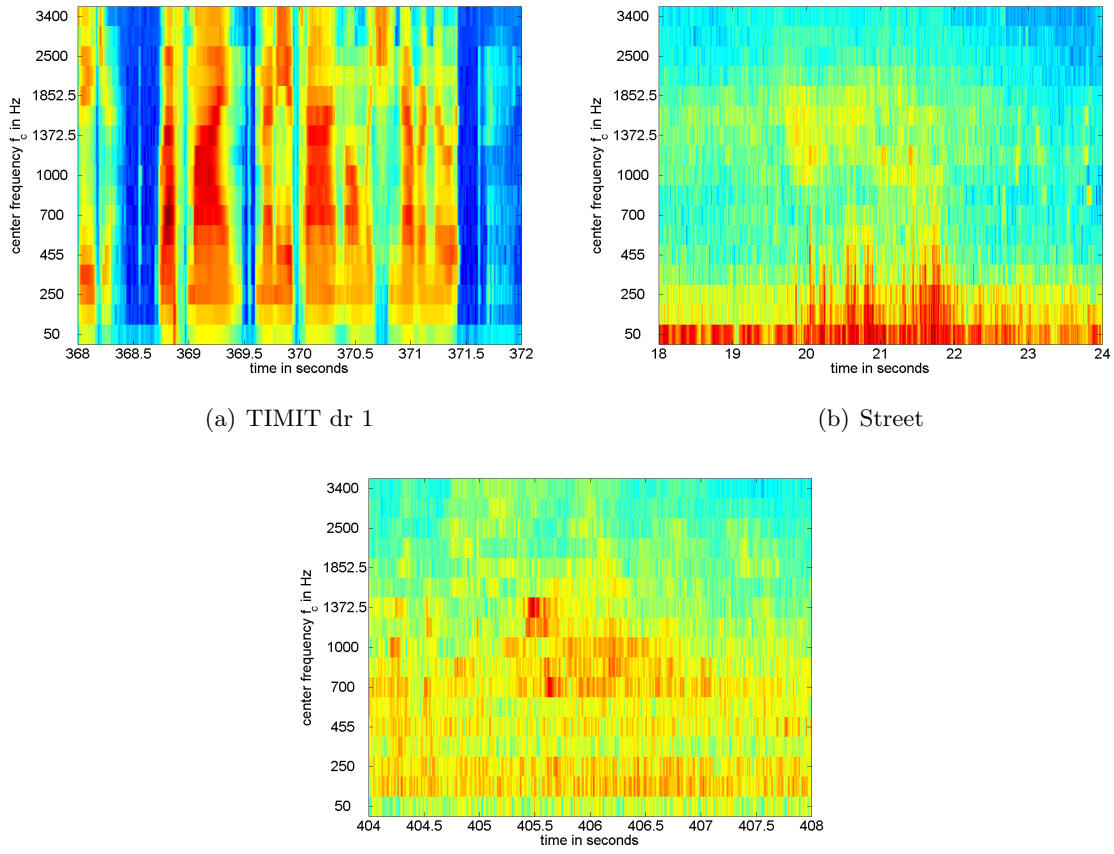
3 Experiments and Results

3.1 Sound Data

The data used for the experiments reported here was selected from a larger database of speech and environmental sounds that has in part been recorded under DIRAC and that in part contains recordings from several other datasources, including speech and noise databases and commercial audio CDs. Data sources are summarized in Fig. 1. Data used for training was balanced (same prior probability for each class) with a length of about 5 minutes. Data used for testing was also balanced with length also about 5 minutes. Additional testing data was collected from the NOISEX sound database and a commercial audio CD. Example spectrogram sections and time-averaged amplitude modulation spectrograms are shown in Fig. 1 and 2, respectively.

Speech class	
Dialect region 1 (dr1) trainset and testset	TIMIT
Dialect region 2 (dr2) testset	TIMIT
Street class	
Roadtraffic 3m distance from road	DIRAC
Roadtraffic close to road	DIRAC
Pedestrian Zone class	
Downtown, near Shop 1	DIRAC
Downtown, near Shop 2	DIRAC
Pedestrians in City Center	CD “1111 Geräusche”
Shopping Mall	CD “1111 Geräusche”
Additional test data	
volvo	NOISEX
factory1	NOISEX
babble	NOISEX
speech	Audiological test data CD

Table 1: Sound data employed for the experiments. Speech data was taken from training and test set portions of the TIMIT database of continuous English speech which contains speech from different dialect regions. Street class training and test set was taken from a recording conducted within DIRAC, containing street sounds recorded near a busy street with car traffic. Pedestrian zone class data was collected from a CD called “1111 Geräusche” (“1111 Sounds”) and from DIRAC recordings. Additional test data was taken from the NOISEX database and from an audio CD for audiological testing.

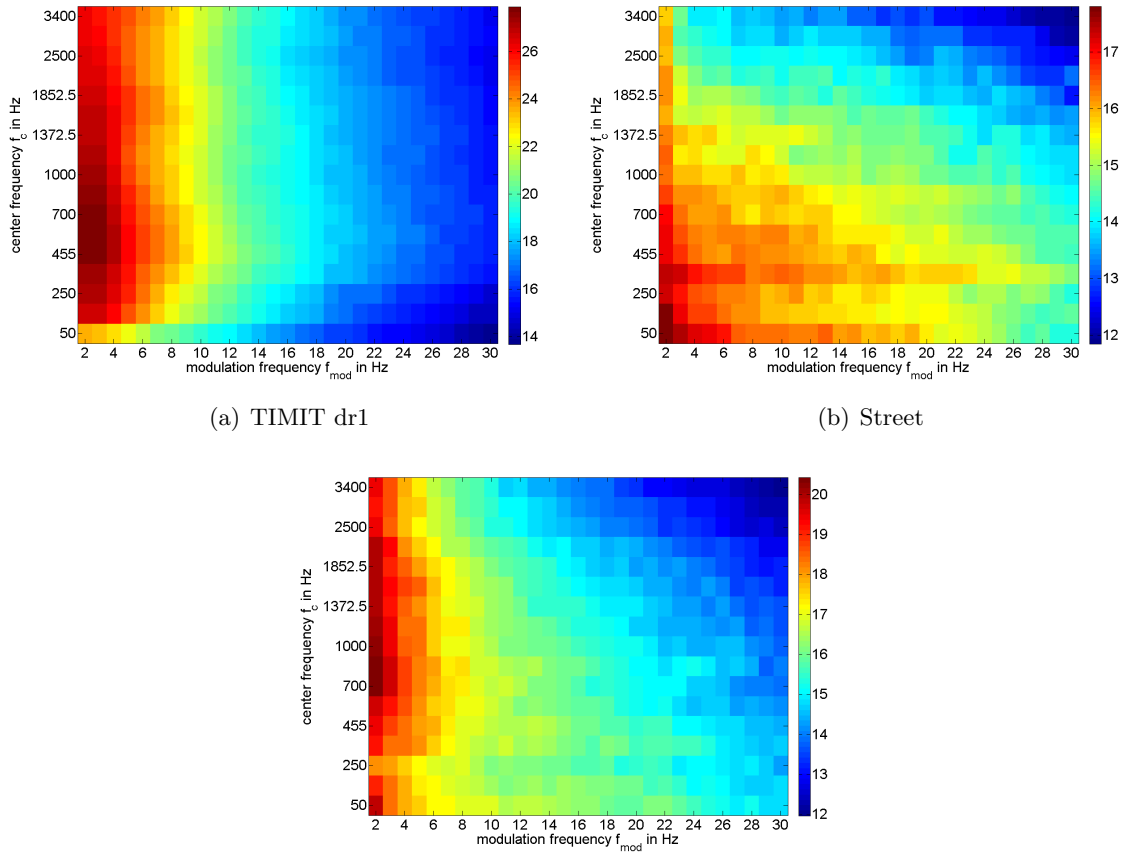


(a) TIMIT dr 1

(b) Street

(c) Pedestrian Zone

Figure 1: Example spectrograms for each of the three classes: (a) Speech, (b) Street and (c) Pedestrian Zone.



(a) TIMIT dr1

(b) Street

(c) Pedestrian Zone

Figure 2: Time-averaged amplitude modulation spectrograms computed from the training data of the three classes: (a) Speech, (b) Street and (c) Pedestrian Zone. Abscissa denotes modulation frequency f_m, ordinate denotes center frequency f_c.

3.2 Pilot Experiments

3.2.1 Model Selection: Variation of Kernel and Parameters

Several pilot experiments were conducted to determine the influence a variation of kernel and classification parameters of the support vector machine classifier has on the AMS based sound classification. The three standard SVM kernels (linear, polynomial and radial basis function) were compared with different margin and kernel parameters with respect to their classification performance on full AMS patterns of (speech-like) babble noise versus factory noise. Performance was evaluated using five-fold cross-validation.

Linear kernel The margin parameter C was varied over a range from 2^{-25} to 2^{15}. Results indicated an influence only for extremely small values of C, as shown in table 2.

Polynomial kernel Margin parameter C and inner-product parameter γ were varied as specified in table 3. The degree of the polynomial was initially chosen as 2. For a wide parameter range the influence on the classification accuracy very small. Increasing the degree of the polynomial resulted in decreasing cross-validation accuracy for degrees of 8 or higher.

Radial Basis Function (RBF) Kernel Parameter varied for the RBF kernel were C (margin parameter) and γ (width parameter of the Gaussian), the corresponding accuracies are shown in table 4. The Gaussian kernel achieves perfect cross-validation performance for one particular parameter setting, however, its performance tends to show larger variance with respect to choice of parameter values.

Cost	Cross validation accuracy
2^{-25}	94.8
2^{-23}	96.7
2^{-21}	98.5
2^{-19}	99.0
2^{-17}	99.5
2^{-15}	99.5
2^{-13}	99.5
2^{0}	99.5
2^{1}	99.5
2^{3}	99.5
2^{5}	99.5
2^{7}	99.5
2^{9}	99.5
2^{11}	99.5
2^{13}	99.5
2^{15}	99.5

Table 2: Model selection with linear kernel. The cross validation accuracy in % is computed with the varying model parameter C.

C / γ	2^{-5}	2^{-3}	2^{-1}	2^{0}	2^{1}	2^{3}	2^{5}	2^{7}
2^{-25}	96.6	97.8	98.5577	98.7	98.9	99.4	99.7	99.8
2^{-23}	95.5	95.5	97.1154	98.4	98.7	99.0	99.5	99.7
2^{-21}	97.3	97.8	98.7179	98.9	99.0	99.5	99.7	99.8
2^{-19}	98.4	98.6	99.3590	99.4	99.5	99.7	99.7	99.7
2^{-17}	99.0	99.2	99.6795	99.7	99.7	99.7	99.7	99.7
2^{-15}	99.5	99.7	99.6795	99.7	99.7	99.7	99.7	99.7
2^{-13}	99.7	99.7	99.6795	99.7	99.7	99.7	99.7	99.7
2^{-11}	99.7	99.7	99.6795	99.7	99.7	99.7	99.7	99.7

Table 3: Model selection for polynomial kernel, cross-validation accuracy in percent in dependence on parameters C and γ.

C / γ	2^{-5}	2^{-3}	2^{-1}	2^{0}	2^{1}	2^{3}	2^{5}	2^{7}
2^{-21}	95.2	96.8	98.7	98.9	99.2	99.4	99.7	100.0
2^{-19}	97.1	98.7	99.2	99.5	99.4	99.7	99.8	99.8
2^{-17}	98.7	99.2	99.5	99.7	99.7	99.8	99.8	99.8
2^{-15}	98.6	99.0	99.5	99.7	99.7	99.7	99.7	99.7
2^{-13}	96.3	96.3	98.7	98.9	98.9	98.9	98.9	98.9
2^{-11}	52.1	52.1	52.1	76.4	76.9	76.9	76.9	76.9

Table 4: Model selection for RBF kernel, cross-validation accuracy in percent in dependence on parameters C and γ.

The preliminary experiments do not seem to suggest that the use of non-linear kernels is necessary since already the linear kernel shows very good classification accuracy. The possible gain of non-linear kernels brings about more parameters whose value has to be optimized, which may impact generalization to new data, and for poor parameter choices the performance drops below that of the linear classifier. Results of the preliminary experiments can only provide indications regarding the final parameter and kernel choice. The subsequent results reported below support these indications, since very good performance is already achieved with the standard parameter setting of the linear SVM.

3.2.2 Feature Selection of entire modulation frequency bands

In this experiment, feature selection is performed for entire "vertical slices" of the amplitude modulation spectrogram, i.e., classification is based on all center-frequency bands (all f_c) of the selected modulation frequency (f_m) bands. Selection is performed using the speech and street classes, cf. table 1.

Table 5 shows that 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 (17 center frequencies per modulation frequency) while at the same

iteration	modulation frequency bands f_m	CV accuracy
1	3 Hz	99,4%
2	3, 4 Hz	99,3%
3	3, 4, 26 Hz	99,5%
4	3, 4, 26, 25 Hz	99,6%
5	previous bands & 9 Hz	99,7%
6	previous bands & 14 Hz	99,7%
7	previous bands & 28 Hz	99,7%
8	previous bands & 13 Hz	99,7%
9	previous bands & 20 Hz	99,8%
	entire AMS pattern	99,7%

Table 5: Feature selection with complete modulation frequency (f_m) band as search parameter. The most salient f_m-band selected per iteration of the feature selection algorithm (including the selected in previous iterations) and the corresponding cross validation accuracy CV_{Acc} are shown.

time the accuracy remains essentially unchanged compared to using the entire AMS pattern for classification (which yields $CV_{Acc} = 99.7\%$).
The experiment shows that it is possible to reduce the number of features while the cross validation accuracy still reaches values near 100%, implying that morecompact feature subset with still very good classification performance should exist. Hence, the feature selection approach in the remaining sections is to select individual combinations of center- and modulation frequencies, i.e., sets of (f_c, f_m)-bins that maximize classification accuracy.

3.3 Training with Clean Speech Data

This section investigates the performance of the proposed classification scheme for different numbers of features when using clean speech for training. As mentioned above, a single feature is defined as an (f_c, f_m) pair in the spectral/modulation-spectral plane of the amplitude modulation spectrogram. The preliminary experiments described above led us to use a linear SVM with margin parameter $C = 1$
Classifier training and feature selection are performed on the tasks of discriminating clean speech vs. street traffic noise, clean speech vs. sound from a pedestrian zone and clean speech vs. a combination of both, respectively. Evaluation is performed with the testing portion of the respective speech and noise data sets, and noise signals from different recordings than trained on in order to test generalization performance. Evaluation is also performed for speech mixed with noise signals at different signal-to-noise ratios (SNRs).

3.3.1 Training vs. Street Noise

Learning to discriminate clean speech from street noise and selecting features using SFFS produces the results shown in Fig. 3, demonstrating that perfect cross-validation performance is obtained with as few as 9 points in the center-frequency/modulation-frequency plane. The most recently added feature for each feature subset size is indicated in Fig. 3 (right). The single most important feature is the 455 Hz center-frequency band and 3 Hz modulation-frequency band feature, which is also located near the modulation energy peak in the AMS pattern. The first 9 features are all located in the modulation frequency range from 2 to 12 Hz,

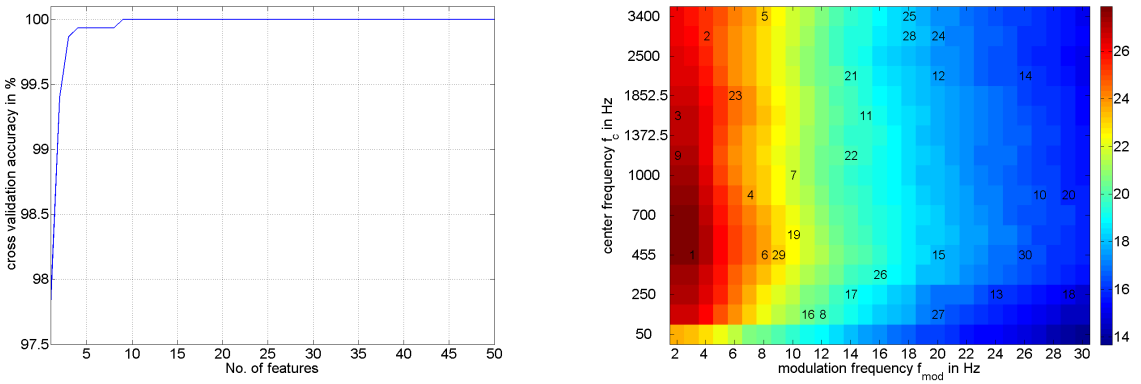


Figure 3: Training on clean speech vs. street noise using SFFS. Left: Cross-validation accuracy as function of number of features. Right: Center-frequency/modulation-frequency positions of first 50 selected features, superimposed on the averaged AMS pattern of speech training data. Numbers indicate feature number corresponding to left plot. Note that only positions of the first 9 features are relevant as accuracy remains at 100% thereafter, leading to near random scattering.

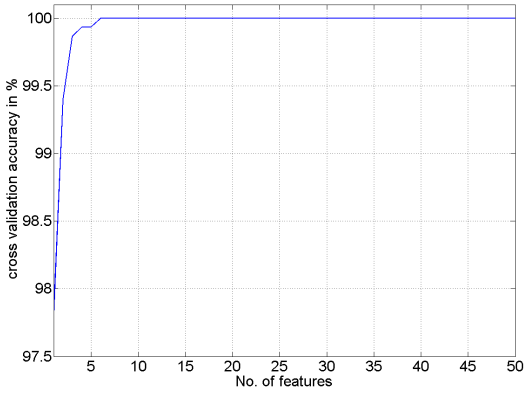


Figure 4: Same experiment as Fig. 3 but selection performed with SFS. Cross-validation accuracy as function of number of features.

with 7 of them in the 2 to 8 Hz range, covering a wide range of center frequencies. These results are compatible with the literature results outlined above. The features for selection iterations beyond 9 appear to be randomly scattered over the entire AMS pattern since they are essentially determined at random as in each subsequent feature selection iteration many additional features will result in a performance of 100% (except those few cases where an additional feature decreases performance).

For comparison, results of the same experiment when using SFS for features selection are displayed in Fig. 4. Behavior of the cross-validation accuracy curve is very similar. Perfect cross-validation accuracy happens to be obtained 3 iterations earlier than with SFFS, which appears to be a result of the random selection of features if more than one feature produce the same accuracy during any iteration of the selection algorithm.

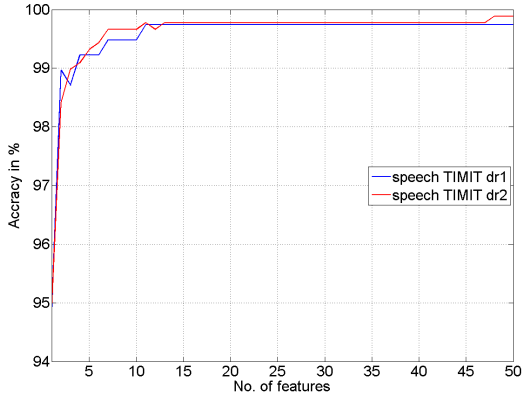


Figure 5: Classification of clean speech vs. street noise: Accuracy on speech test data (not used for cross-validation). "TIMIT dr1" corresponds to the same dialect used for training, "TIMIT dr2" is a different dialect.

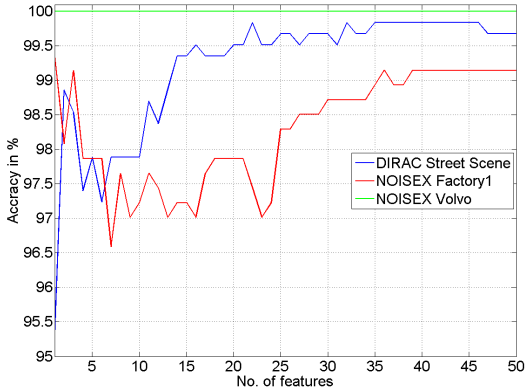


Figure 6: Classification of clean speech vs. street noise: Accuracy on noise signals not included in training data. "DIRAC street scene": street noise recorded at a different location than training data, "NOISEX Factory 1": machine noise with strong non-stationary component, "NOISEX Volvo": very stationary noise floor inside driving car.

Generalization to new data Generalization capabilities of the selected features have been tested using data not included in the training set (i.e., data that has not been used to compute cross-validation accuracy and feature selection), displayed in Fig. 5 for new speech data and Fig. 6 for new non-speech data. Performance of new data tends to increase with number of selected features, albeit slower than on the training data set. Speech drawn from a different ensemble (different dialect region) is classified about as good as test data from the same ensemble (same dialect region as in training), indicating that sufficiently universal speech properties are picked up by the selected features. Performance on new noise recordings as expected depends on the type of noise source, with stationary engine noise reliably being identified as more similar to street noise than to speech. New street noise is identified very well with about 40 features. Non-stationary machine noise from a factory classified less reliable, but still with an accuracy of over 90% when using 40 features.
Pauses in a speech signal are sometimes classified as non-speech if they exceed the length of

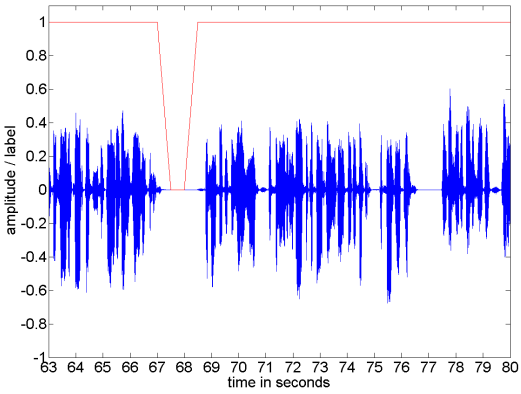
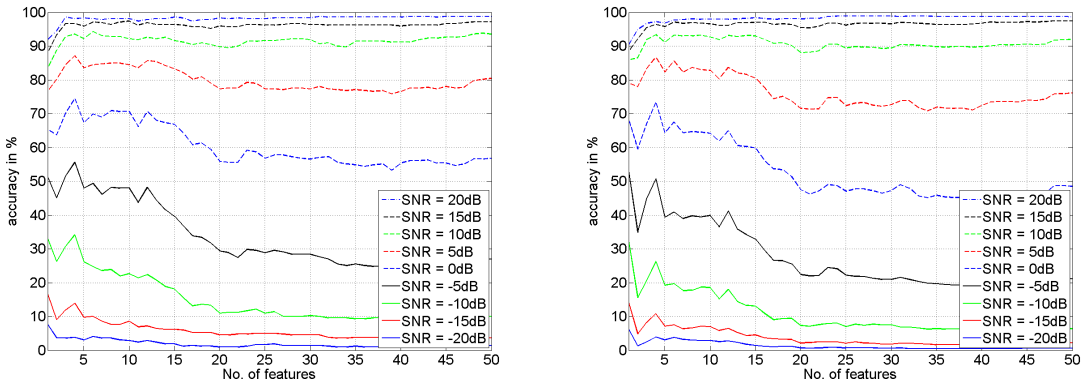


Figure 7: Classification of clean speech vs. street noise: Example of applying the trained classifier to a speech signal from a different data source than training speech data. Speech is identified almost perfectly; only in speech pauses with length larger than the feature analysis window (1 s) are some pauses classified as non-speech (also depending on the precise position of the analysis window with respect to the speech pause).

the analysis window (1 s) and the analysis window falls into the gap, cf. Fig. 7.

Test on non-clean speech data Performance for classification of speech vs. street noise has also been tested by embedding speech in a background of street noise at different signal-to-noise-ratios (SNRs) and applying the same classifier that has been trained on clean speech vs. street noise data. Fig. 8 demonstrates that, as expected, speech at high ("good") SNR tends to be classified at speech, whereas speech embedded in a dominating noise signal tends to be classified as noise. For a feature subset size of about 40, an SNR of 0 dB produces roughly the same probability of classification for either class which reflects prior probabilities learned from the balanced training set, cf. also Fig. 9. For small feature numbers there appears to exist a bias towards classifying the signal as speech, reinforcing the earlier observation that very small feature numbers do not result in best generalization, but rather modest numbers of about 40 features should be used. Testing data from the same dialect region ("dr1") at low SNRs at 40 features tends to give an about 5 % point higher likelyhood for being classified as speech as compared to test data from a different dialect region ("dr2"). This may indicate statistical differences between the two data ensembles that have been picked up by the algorithm and that might be related to the different dialects. The present experiments, however, are not sufficient to rule out other possible explanations not related to the speech dialect.



(a) TIMIT dialect region 1

(b) TIMIT dialect region 2

Figure 8: Training on clean speech vs. street noise, testing on speech mixed with street noise at different signal-to-noise ratios (SNRs). Speech data from two dialect regions.

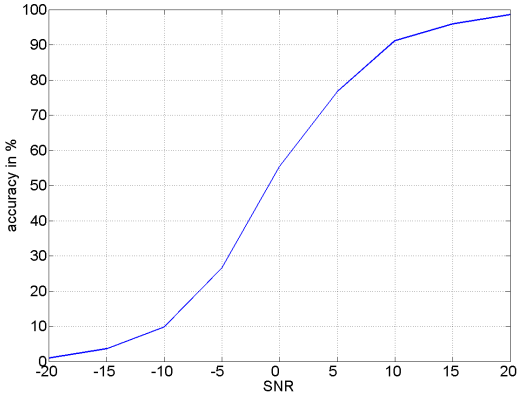


Figure 9: Training on clean speech vs. street noise, testing on speech mixed with street noise at different signal-to-noise ratios (SNRs). Obtained classification accuracies using 40 features on TIMIT dialect region 1 test data, cf. also to Fig. 8.

3.3.2 Training vs. Pedestrian Zone Noise

Classification of clean speech versus ambient noise recorded in a busy pedestrian zone area has been conducted with experiments in analogy to those reported in the previous section for speech vs. street noise classification. Results are displayed in Fig. 10 to 13. The averaged modulation spectrogram of pedestrian zone noise is more similar to speech than that of street noise is to speech (cf. Fig. 2), which is expected as pedestrian zone noise contains speech babble as a significant component, but no engine noise at all. Hence, it might be expected that classification of speech vs. pedestrian zone noise is in principle harder than speech vs. street noise. Results for cross-validation accuracy (Fig. 10) and speech test data (Fig. 11) confirm this. Classification of different noise signals (Fig. 12) is biased towards detecting non-speech compared to the previous section. Testing of the trained classifier with speech mixed with pedestrian zone noise at different SNRs (Fig. 13) further illustrates this with a strong bias towards detecting pedestrian zone noise instead of speech. Clearly, slightly degraded speech has to be subsumed into the noise class since it may likely have been just a background babble in the pedestrian zone scene.

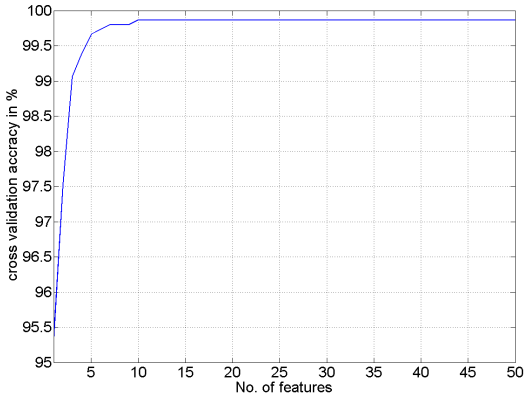


Figure 10: Classification of clean speech vs. pedestrian zone noise: Cross-validation accuracy as function of number of features.

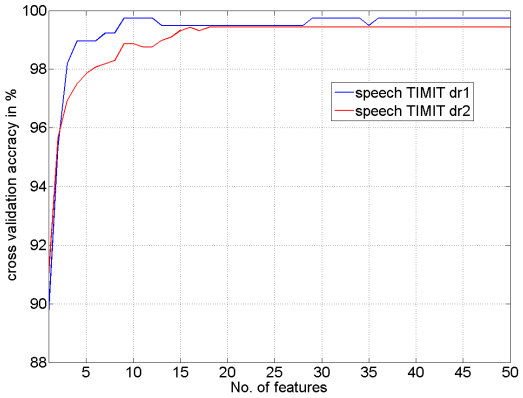


Figure 11: Classification of clean speech vs. pedestrian zone noise: Accuracy on speech test data.

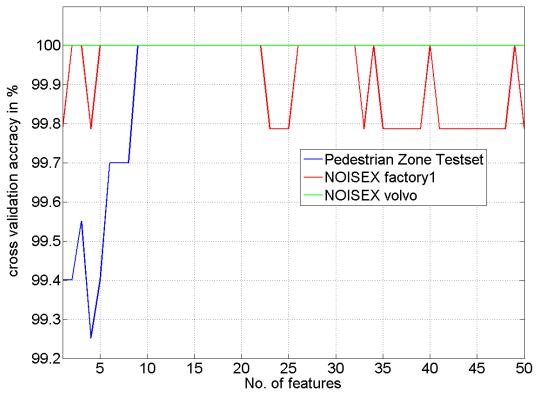
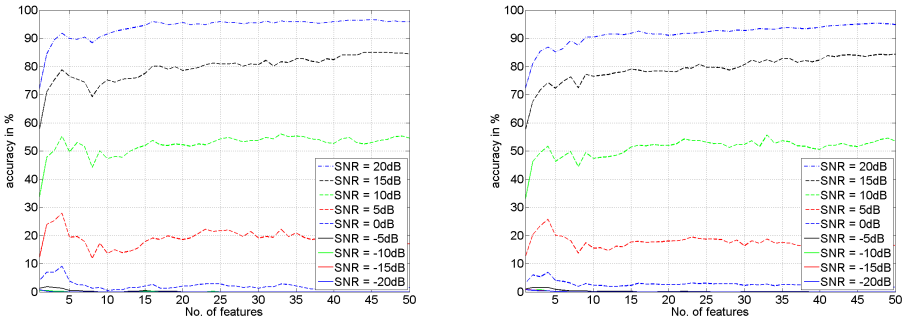


Figure 12: Classification of clean speech vs. pedestrian zone noise: Accuracy on noise signals not included in training data.



(a) TIMIT dialect region 1 (b) TIMIT dialect region 2

Figure 13: Training on clean speech vs. pedestrian zone noise, testing on speech mixed with pedestrian zone noise at different signal-to-noise ratios (SNRs). Speech data from two dialect regions.

3.3.3 Training vs. Street & Pedestrian Zone Noise

Experiments of the previous sections were repeated with training discrimination of clean speech against street and pedestrian zone noise, i.e., the non-speech training class was composed of an equal amount of street noise and pedestrian zone noise AMS patterns. Results are highlighted in Fig. 14 and 15 and are similar to the ones obtained from training on speech vs. pedestrian zone noise.

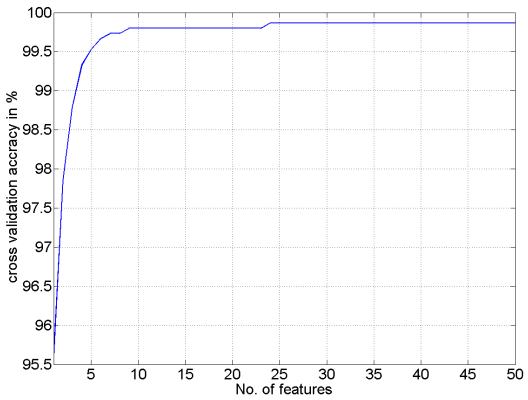


Figure 14: Classification of clean speech vs. street & pedestrian zone noise. Cross-validation accuracy as function of number of features.

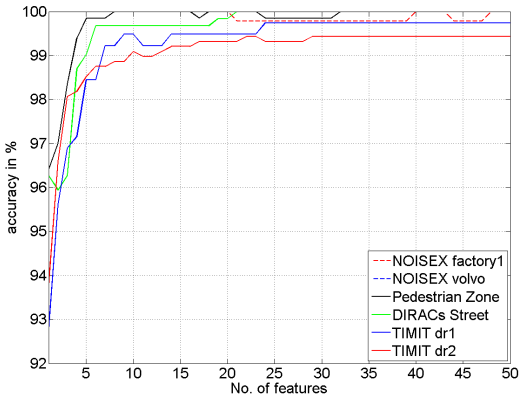


Figure 15: Classification of clean speech vs. street & pedestrian zone noise: Accuracy on various speech and non-speech test data (not used for cross-validation).

3.4 Training with Speech Data embedded in Background Noise

The previous sections have shown that speech and non-speech acoustic scenes can be discriminated very well based on a modest number of modulation features. As has also been shown above, the learned classifiers are not appropriate for detecting speech in a background of the trained non-speech signal since at a certain threshold the algorithm will "consider" the presented input as more similar to the non-speech signal even if some speech component is still present in it.

This section investigates detection of speech embedded in a realistic background such as the street scene and pedestrian zone scene signals. Classifier training is performed for discrimination between speech embedded in the background signal and the pure background signal without the presence of speech. Two scenarios are considered: Speech embedded in street scene noise and speech embedded in in pedestrian zone noise. For each scenario, speech in background is discriminated from the pure background in two ways. (a) Signal-to-noise ratio of speech in the background fixed at a single SNR. This is done for 9 fixed SNR values from -20 dB to 20 dB in 5 dB steps, resulting in one trained classifier per SNR value, cf. Sec. 3.4.1 and 4.0.3. (b) Speech embedded in the background with a variable SNR ranging from 0 dB to 20 dB. A single classifier is obtained that has no knowledge of the SNR of each observation, cf. Sec. 3.4.2 and 4.0.4.

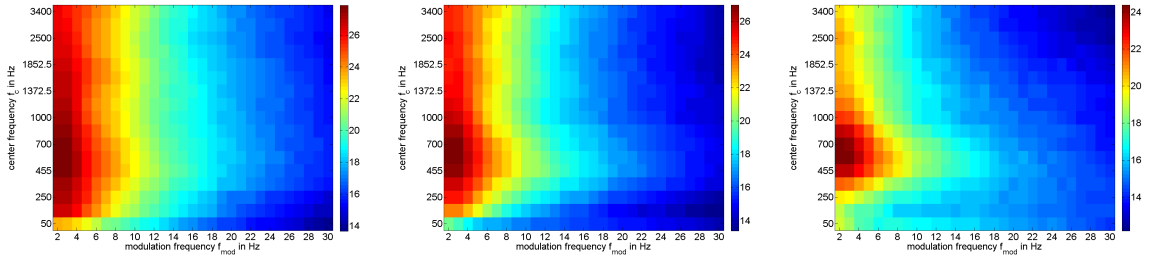
3.4.1 Speech in street noise background at different fixed SNRs

Speech was embedded in a street background at different fixed SNRs ranging from -20 dB to 20 dB in 5 dB steps. The resulting averaged AMS patterns of the training data for three example SNRs (and clean speech and "clean" noise) are displayed in Fig. 16.

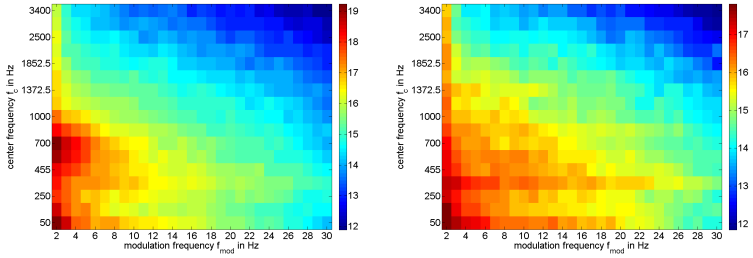
Cross-validation accuracy as a function of number of selected features and training SNR is displayed in Fig. 17, showing that discrimination at SNR values above 5 dB degrades only moderately, exceeds 90% correct at an SNR of -10 dB and remains well above chance level for SNR -20 dB.

Center-frequency/modulation-frequency positions of selected features are shown in Fig. 18, corroborating the importance of modulations below 8 Hz for detecting speech. Compared to the results obtained with clean speech vs. street noise, selected features have narrowed down in their extent on the center-frequency dimension, focussing much closer around the center-frequency/modulation-frequency peak of average modulation energy.

Characteristics of correct positive speech detections vs. false positive speech detections, similar to threshold-derived ROC (receiver-operating-characteristic) curves, are obtained by evaluating the various classifiers trained at different fixed training SNRs with respect to their performance at different fixed test SNRs. The resulting ROC-like curves are plotted in Fig. 19 and the performance data in Fig. 20. The results demonstrate that training at very high or very low SNRs is not desired when test data from different SNR regimes is used. Training in the range of about -5 dB to 10 dB may yield a better trade-off between correct- and false-positives for many applications.



(a) clean speech (b) SNR = 20dB (c) SNR = 0dB



(d) SNR = -20dB (e) "clean" street noise

Figure 16: Averaged AMS patterns computed from speech (TIMIT, dialect region 1) mixed with street noise at different fixed SNRs.

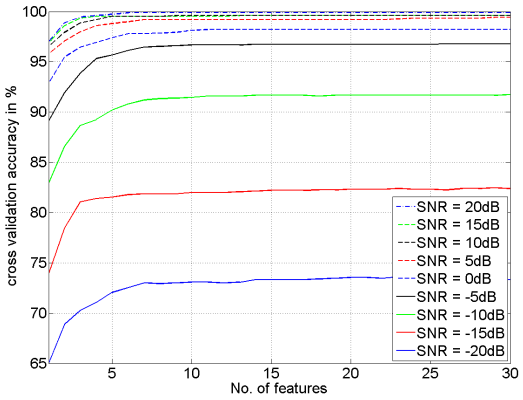
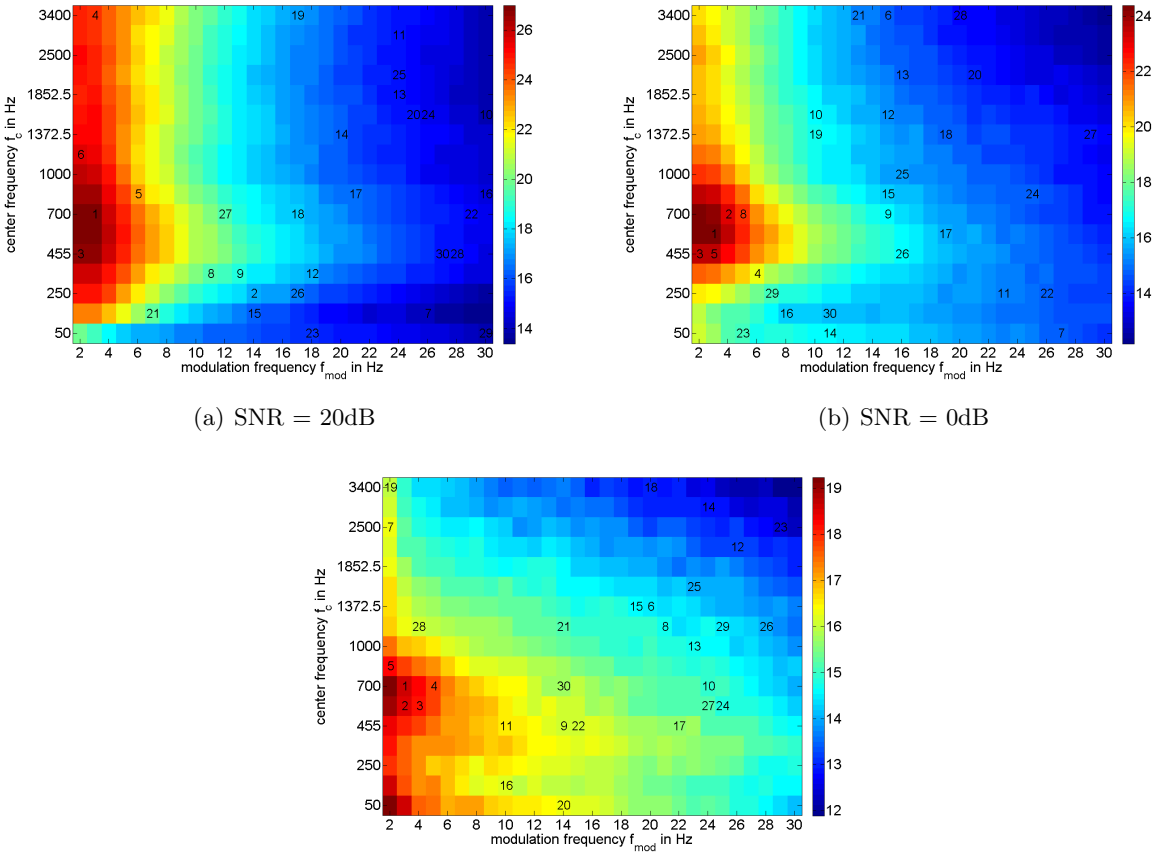


Figure 17: Training on speech embedded in street background at different fixed SNRs vs. street noise with SFFS feature selection. Cross-validation accuracy as function of number of features for different SNR levels.

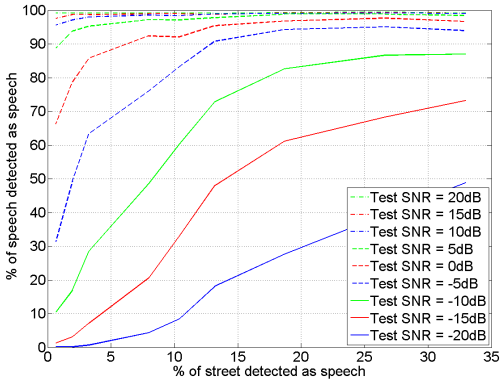


(a) SNR = 20dB

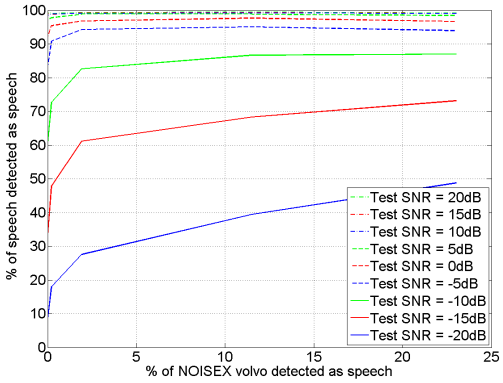
(b) SNR = 0dB

(c) SNR = -20dB

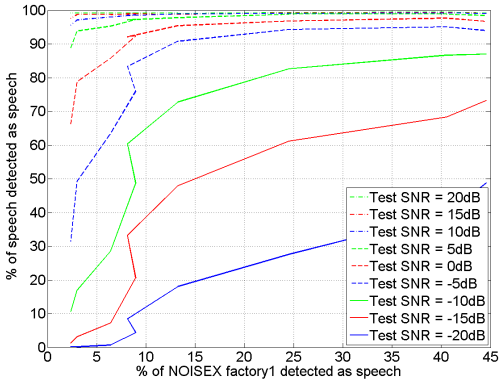
Figure 18: Training on speech embedded in street background at different fixed SNRs vs. street noise. Center-frequency/modulation-frequency positions of 50 selected features, superimposed on the averaged AMS pattern of speech training data for different SNRs. Note that the features selected before cross-validation accuracy saturates (after about 5 to 8 features) are relevant, whereas subsequent features are largely scattered at random.



(a) Speech in street noise

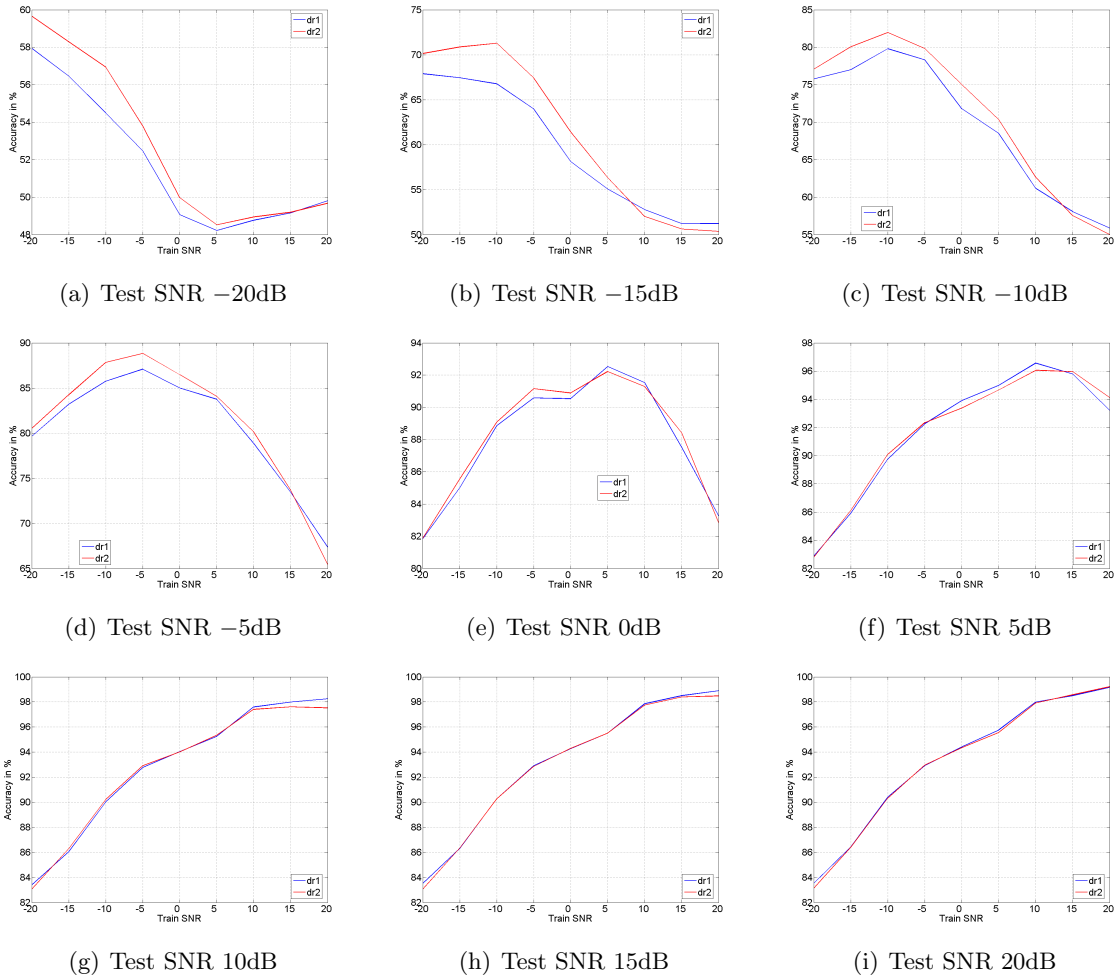


(b) Speech in volvo



(c) Speech in factory1

Figure 19: Training on speech embedded in street background at different fixed SNRs vs. street noise. "Receiver-operating-characteristic (ROC)"-like curves of true-positives vs. false-positives rates (in percent) for testing on different test data at different test SNRs. Note that the parameter varied along each curve is the fixed training SNR, ranging from 20 dB at the curves' lower left end to −20 dB at the upper right.



(a) Test SNR −20dB (b) Test SNR −15dB (c) Test SNR −10dB

(d) Test SNR −5dB (e) Test SNR 0dB (f) Test SNR 5dB

(g) Test SNR 10dB (h) Test SNR 15dB (i) Test SNR 20dB

Figure 20: Training on speech embedded in street background at different fixed SNRs vs. street noise. Accuracy shown for testing on speech test data (dialect regions 1 and 2, respectively) embedded in street noise background at different fixed SNRs vs. street noise test data.

3.4.2 Speech in street noise background at variable SNRs

Motivated by the results of the previous section, a single classifier was trained on discrimination between speech embedded in a street noise background at variable SNRs between 0 dB and 20 dB and "pure" street noise background. Resulting cross-validation performance and performance on test data is shown in Fig. 21. Since no parameters could be varied as in the previous section, there is no corresponding ROC curve plot. The performance on test data with this experiment is slightly inferior to that of a classifier trained at a fixed SNR of 5 dB in the previous section. Hence, more elaborate strategies may be needed to outperform the simple noisy-training strategy of the previous section.

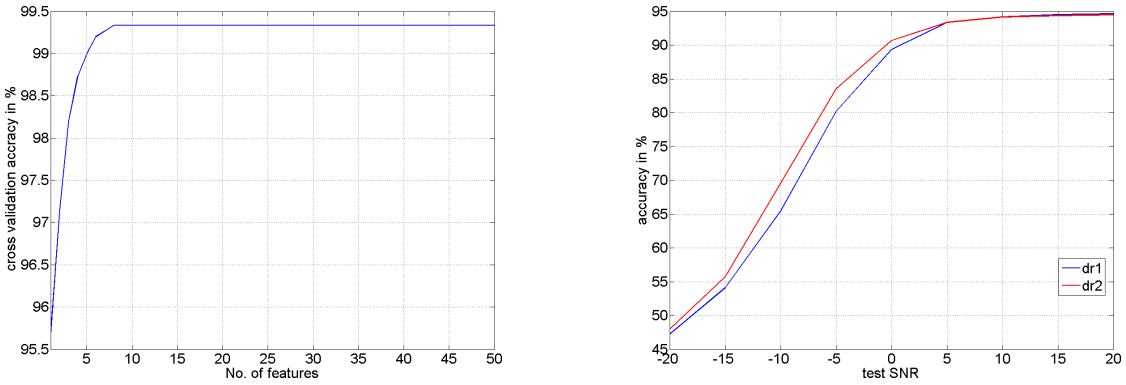
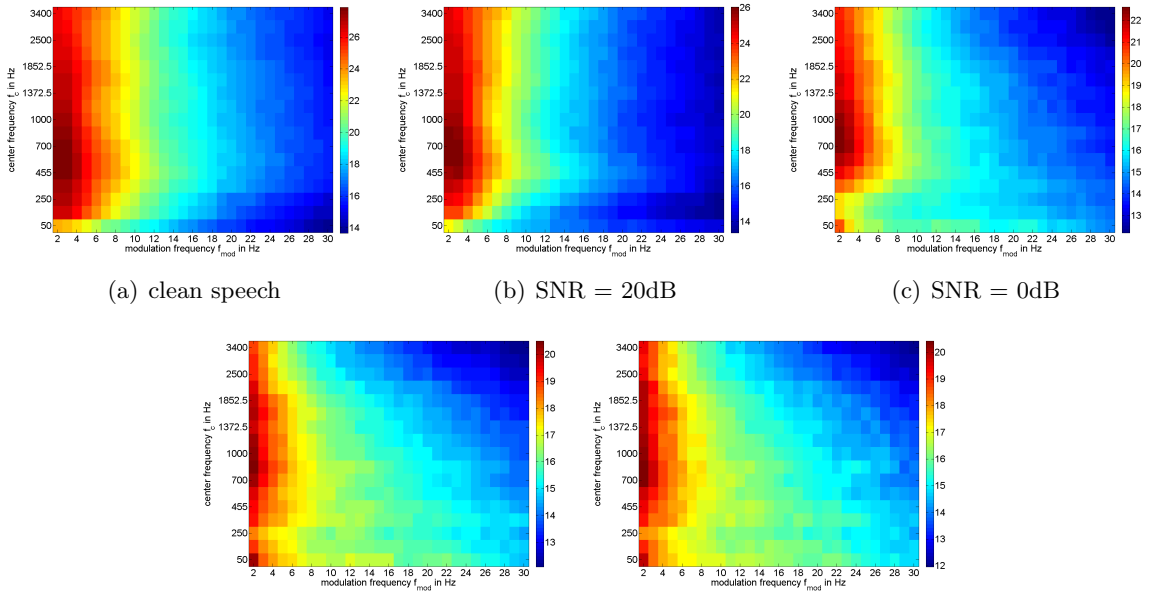


Figure 21: Training on speech embedded in street background at variable SNRs from 0 dB to 20 dB vs. street noise. Left: Cross-validation accuracy as function of number of features. Right: Test on speech test data (dialect region 1 and 2, respectively) embedded in street background at different fixed test SNRs vs. street noise.

4 Conclusion

4.0.3 Speech in pedestrian zone noise background at different fixed SNRs

The experiments of the previous two sections have been repeated for speech embedded in a background of pedestrian zone noise. Speech and pedestrian zone noise are more similar than speech and street noise since pedestrian zone noise is to a significant extent composed of unintelligible babble from persons, a fact that is reflected in the averaged AMS patterns of speech mixed with pedestrian zone noise at different SNRs, as displayed in Fig. 22. Discrimination between speech in pedestrian zone background and pure pedestrian zone noise are therefore harder than in the case of street noise. This is reflected in the cross-validation accuracy for different numbers of features and different SNRs displayed in Fig. 23, and in the performance of the trained SNR-specific classifiers on test data with different SNRs, cf. Fig. 24.



(a) clean speech (b) SNR = 20dB (c) SNR = 0dB

(d) SNR = -20dB (e) “clean” pedestrian zone noise

Figure 22: Averaged AMS patterns computed from speech (TIMIT, dialect region 1) mixed with pedestrian zone noise at different fixed SNRs.

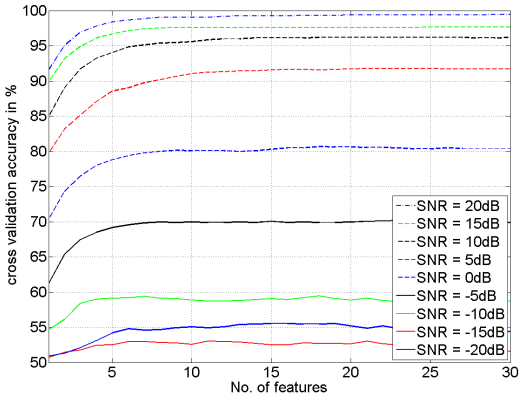
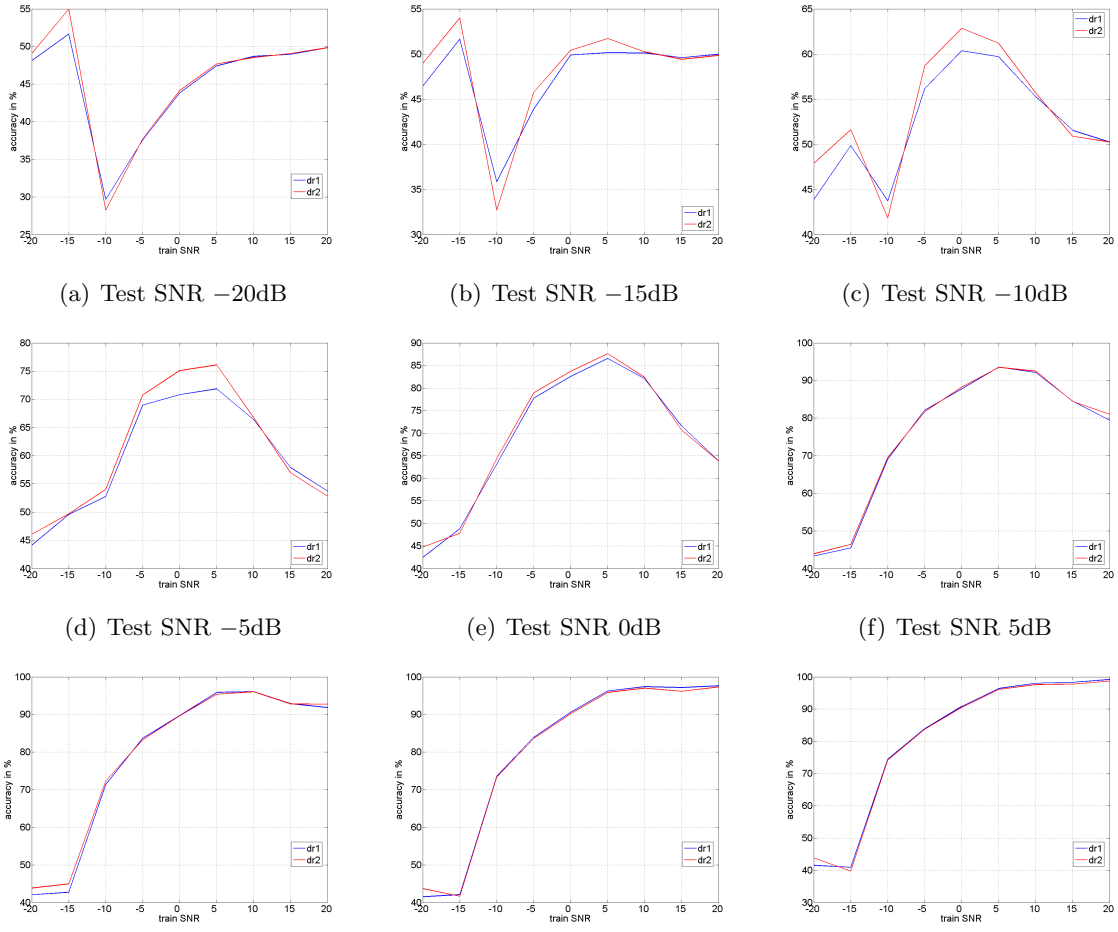


Figure 23: Training on speech embedded in pedestrian zone background at different fixed SNRs vs. pedestrian zone noise. Cross-validation accuracy as function of number of features for different SNR levels.



(a) Test SNR −20dB (b) Test SNR −15dB (c) Test SNR −10dB

(d) Test SNR −5dB (e) Test SNR 0dB (f) Test SNR 5dB

(g) Test SNR 10dB (h) Test SNR 15dB (i) Test SNR 20dB

Figure 24: Training on speech embedded in pedestrian zone background at different fixed SNRs vs. pedestrian zone noise. Accuracy shown for testing on speech test data (dialect regions 1 and 2, respectively) embedded in pedestrian zone noise background at different fixed test SNRs vs. pedestrian zone noise test data.

4.0.4 Speech in pedestrian zone noise background at variable SNRs

In analogy to the experiments for speech in street noise background (Sec. 3.4.2), a single classifier has been trained to discriminate between speech in pedestrian zone background at variable SNRs ranging from 0 dB to 20 dB and pure pedestrian zone noise. The cross-validation performance and the performance on test data with different test SNRs are displayed in Fig. 25. The result reflects the previous section's result that pedestrian zone noise is more difficult than street noise with respect to embedded speech detection. Performance is below the levels obtained in the previous section, indicating that detection of speech in pedestrian zone background is particularly hard when the expected SNR is unknown.

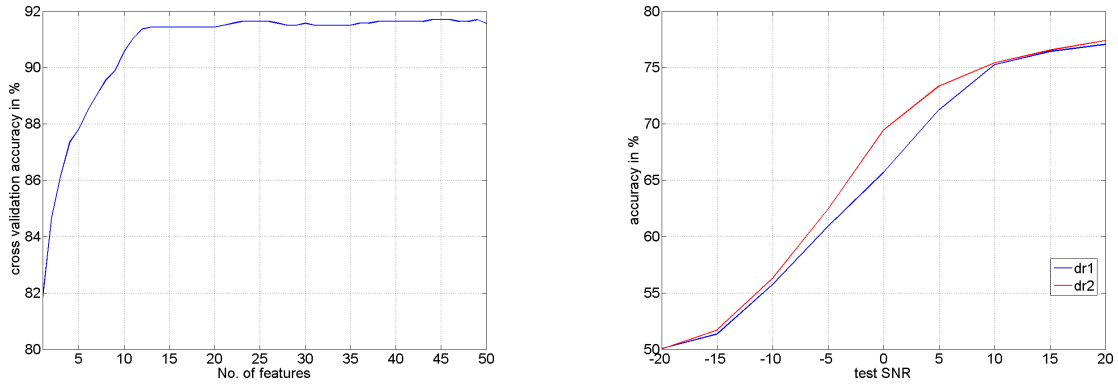


Figure 25: Training on speech embedded in pedestrian zone background at variable SNRs from 0 dB to 20 dB vs. pedestrian zone noise. Left: Cross-validation accuracy as function of number of features. Right: Test on speech test data (dialect region 1 and 2, respectively) embedded in pedestrian zone background at different (fixed) test SNRs vs. pedestrian zone noise.

5 Conclusion

The present work has demonstrated that discrimination between speech and non-speech sounds can be performed with high accuracy based on the amplitude modulation spectrogram (AMS). Detection of speech embedded in a background of non-speech sounds has also been performed based on AMS patterns. The accuracy obtained depends on the signal-to-noise-ratio of speech in its background and on the similarity of the non-speech sounds and speech. If both classes are sufficiently different, detection can be performed with high accuracy even for SNRs of -10 dB. It degrades for recognition in non-speech sounds that bear similarities with the speech signal class, such as pedestrian zone noise. Feature selection was used to find the features leading to highest classification accuracy. These features have been found to be located in a modulation frequency range from about 2 Hz to about 10 Hz, a range that is known in the literature to be highly relevant for speech processing.

References

[1] Büchler, M., Allegro, S., Launer, S. and Dillier, N.: Sound classification in hearing aids inspired by auditory scene analysis. EURASIP Journal on Applied Signal Processing, 18, pp. 2991–3002, 2005

[2] 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`

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

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

[5] 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

[6] Kanedera, N., Arai, T., Hermansky, H., and Pavel, M.: On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Communication, 28, pp. 43–55, 1999

[7] 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

[8] Mesgarani, N.; Shamma, S. A.; Slaney, M.: Speech discrimination based on multiscale spectro-temporal modulations. Proc. ICASSP, 2004

[9] Mesgarani, N., Slaney, M., Shamma, S.A.: Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations. IEEE T. Audi. Speech. Lang. P., 14, pp. 920–930, 2006

[10] Nordqvist, P. and Leijon, A.: An efficient robust sound classification algorithm for hearing aids. J. Acoust. Soc. Am., 115, pp. 3033–3041, 2004

[11] Ostendorf, M., Hohmann, V., and Kollmeier, B.: Klassifikation von akustischen Signalen basierend auf der Analyse von Modulationsspektren zur Anwendung in digitalen Hrgerten [Classification of acoustical signals based on the analysis of modulation spectra for the application in digital hearing aids]. Proc. DAGA, pp. 402403, 1998

[12] Pudil, P., Novovicova, J., Kittler, J.: Floating search methods in feature selection. Pattern Recognition Letters, 15, pp. 1119–1125, 1994