Classification of Environmental Sounds for future Hearing Aid Applications

Jürgen Tchorz[1], *Simone Wollermann*[1,2], *Hendrik Husstedt*[2]

[1]*Fachhochschule Lübeck,* [2]*Deutsches Hörgeräte Institut Lübeck*
tchorz@fh-luebeck.de

Abstract: Different acoustic environments require different hearing aid settings to achieve best speech understanding and sound quality. Manual adjustment of hearing aid settings can be annoying. Thus, many hearing aids automatically classify the acoustic environment and switch between different programs accordingly.
The classification approach presented in this study utilizes so-called amplitude modulation spectrogram (AMS) as features, which replicate aspects of sound analysis in the auditory pathway. The AMS patterns represent time intervals of 500 ms each. The classification of the acoustic environment based on these features is implemented with supervised machine learning using a deep neural network. The network is trained on features extracted from several hours of sound from different classes, namely speech, reverberant speech, speech in noise, music, and noise. For testing, a set of sounds taken from other recordings was processed and classified by the neural network. For comparison, these sounds were also automatically classified using hearing aids from five different brands. The results show comparable classification accuracy with amplitude modulation spectrograms and hearing aids, respectively. The time which is needed to classify a situation, however, is much shorter with the amplitude modulation spectrogram-based approach.

1 Introduction

Humans are able to classify audio signals without conscious effort. For us, it is easy to tell whether a certain sound is speech, noise, or music, as long as the sound is not too weak or masked. In addition, human sound classification is quite fast: we can almost instantaneously identify the origin of a certain sound. Also, more complex classification tasks are managed in short time: in a study on music genre recognition, Gjerdingen et al. [1] showed that segments of 250 ms allow for genre recognition well above chance. In several technical applications, an automatic classification of environmental sounds is desired. In hearing aids, for example, such a classification of the acoustical environment is widely used. Depending on the detected situation, different gain and feature settings such as noise suppression or directional microphones are automatically activated to improve speech understanding, sound quality or listening comfort. Typical sound categories are speech in quiet, speech in noise, music or noise.

Environmental sound classification in hearing aids is mostly based on simple features like profile and temporal changes of the frequency spectrum, statistical distribution of signal amplitudes, or analysis of modulation frequencies, and subsequent heuristic classification approaches [2, 3]. Typically, these algorithms need several seconds to detect a change in the acoustic situation and to adjust the hearing aid setting accordingly.

Büchler et al. [4] compared different combinations of features inspired by auditory scene analysis, such as tonality, pitch variance, width of the amplitude histogram, level fluctuation

strength and others with different pattern classifiers (e.g., rule-based, Bayes, two-layer perceptron, and Hidden Markov models). The classifies were trained on four sound classes (speech, speech in noise, noise and music). For each sound of 30-second length, classification was calculated once per second, and the class that occurred most frequently was taken as an output for that sound. Depending on the combination of feature set and pattern classifier, classification rates between about 80 - 90% were achieved for these 30-seconds segments. "Speech in noise" was classified slightly poorer than the other classes, particularly in very low or very high SNR, which was classified as “noise” or “speech,” respectively.

Recently, Pita et al. [5] proposed a computationally efficient classification approach which utilizes features determined from temporal statistics of Mel frequency cepstral coefficients (MFCCs) and a multilayer perceptron with up to 20 hidden layers as classifier.

The features used in this study (Amplitude Modulation Spectrograms, AMS) reflect both spectral and temporal aspects of the input signal. This mimics important aspects of the auditory system, where not only frequency information is represented in a topographical way, but also a gradient of amplitude modulation rate representation can be found. In neurophysiological experiments, several researchers found neurons in the inferior colliculus and auditory cortex of mammals which were tuned to certain modulation frequencies. The “periodotopical” organization of these neurons with respect to different best modulation frequencies was found to be almost orthogonal to the tonotopical organization of neurons with respect to center frequencies. Thus, a two-dimensional “feature set” represents both spectral and temporal properties of the acoustical signal (see Baumann et al. [6] for a review). AMS features have previously been used for SNR estimation [7] and noise suppression [8] in combination with Bayesian or neural network classifiers.

For this study, a deep neural network was trained on AMS features extracted from sound samples of different environments which are relevant for hearing aid users (speech, speech in noise, reverberant speech, noise, and music). Another set of sound samples was used to test the classifier. The results were compared to classification of the same set of sounds with hearing aids from different manufactures.

2 Method

2.1 Sound database

In total, 24 hours of sound were used for training (almost 5 hours for each of the five sound classes). Speech was taken from the Buckeye Corpus of conversational speech [9]. In total, 35 speakers (m/f) were included. The speech material is phonetically labeled which allowed for automated removal of speech pauses, coughs, sighs and other non-speech portions. The noise samples (n=82) included traffic noise, nature sounds, crowds without dominant voices, and machinery. The music samples did not include many different genres but consisted of 103 pop songs from the 1960s. Speech in noise was adjusted to a fixed overall SNR of 5 dB (generated from randomly selected above described speech and noise samples). Reverberant speech was produced by convolving the speech samples with the impulse response of a small church (1.5 s average reverberation time).

The test set included two speech samples (the International Speech Test Signal [10] and a poem read by a male speaker), two noise samples (unmodulated speech-shaped noise and a rock-sieving machine), two classical music samples (solo piano and violin), and mixtures of the above described speech and noise samples at SNRs of 5 and 15 dB, respectively. Reverberant speech was not included in the test samples. The overall length of the test samples was 22 minutes. Only a limited number of test samples was included, as the classification experiments

with the hearing aids could not be automated and were quite time-consuming.

2.2 Feature extraction

Prior to AMS pattern generation, the sound files were downsampled to 12 kHz (16 bit). Then, in brief, FFTs are computed for overlapping 4 ms segments of the signal. The resulting amplitudes in each frequency bin are regarded as envelope signal. The modulation spectra are obtained by computing FFTs in each frequency band across a Hanning-windowed time segment of 500 ms. Appropriate summation of neighboring FFT bins yield 25 frequency channels and 15 modulation frequency channels, i.e., patterns with 375 numbers for each non-overlapping 500 ms segment of the input signal. Details can be found in [7].

2.3 Neural network classifier

For classifying the 500 ms segments, a deep learning recurrent neural network (long short-term memory network, LSTM) with three hidden layers (512 neurons each) was implemented using the Microsoft Cognitive Toolkit (CNTK). The training of the neural network ($\approx$ 170.000 AMS patterns, 200 iterations) took about 25 minutes on a Nvidia GTX 970 graphics processing unit.

2.4 Hearing aid classification

Hearing aids with automatic sound classification switch between different hearing programs, depending on the detected sound environment. A program designed for understanding speech in noisy situations, for example, will activate a directional microphone and a stronger level of noise suppression whenever this situation is detected. The active hearing program, however, cannot be "seen" directly when the hearing aid is used. Thus, the hearing programs which are available for different situations were "marked" prior to the classification experiments: the gain in a narrow frequency range was strongly reduced, so that the output of the hearing aid could be used to check whether the hearing aid activated a certain program, depending on the input.

In total, 5 hearing aids from different brands were examined. The sound samples described above were presented at 65 dB(A) for 120 s each. Only the last 60 s were analyzed to allow for adaptation to the situation. The result of this experiment was the classification rate for each sample, averaged across hearing aids. Not every hearing aid had specific programs for all 4 sound classes (speech, noise, music, and speech in noise, respectively).

3 Results

Table 1 shows the confusion matrix for automatic classification with the proposed AMS-feature approach. The overall classification rate was 76.5%. Noise segments were classified 91% correct. The classification rate for music segments was 80%, although classical music was not included in the training data, which demonstrates that the pattern recognizer generalized to some extend.

The classification accuracy for speech segments was only moderate (77%). Most of the mislabeled segments were classified as "music" (18%). It can be seen from Figure 1 that these errors often occur in speech pauses. In the training data for speech, pauses were not included. One of the two test samples for speech (the poem), however, contained several pauses of a few seconds each. In these speech pauses, the output neuron activity of the music neuron was the highest one. Thus, the classification result was "music" in these portions. An explanation might be that the pauses between the 103 music tracks had *not* been removed prior to training.

Table 1 – Confusion matrix (in per cent) for the AMS-based classifier (500-ms segments)

	Presented				
Classified	Speech	Noise	Reverb.	Music	Speech in Noise
Speech	77	0	-	2	9
Noise	0	91	-	4	27
Reverb.	0	0	-	1	2
Music	18	9	-	80	3
Speech in Noise	4	0	-	13	60

Table 2 – Confusion matrix (in per cent) for the classification with hearing aids (samples)

	Presented			
Classified	Speech	Noise	Music	Speech in Noise
Speech	90	0	30	23
Noise	0	88	0	4
Music	0	0	70	0
Speech in Noise	10	13	0	73

Speech in noise had the poorest classification rate (60%). Most confusions occured with "speech" and "noise". The classifier was also trained to detect reverberant speech, which was not present in the test set. Only very few segments were mislabeled as "reverberant speech".

The confusion matrix for classification with hearing aids is shown in Table 2. In total, ten samples have been presented to all of the five hearing aids. On average, 77.8% of the samples were classified correctly. While speech and noise were identified mostly correct, music and speech in noise were frequently classified as "speech".

4 Discussion

Both the AMS-based classifier and the hearing aids struggled to precisely recognize speech in noise. This is in line with other studies. "Speech in noise" is not really well-defined. In addition, even if the overall SNR is fixed, the *local* SNR in short segments of the signal strongly fluctuates even in stationary noise, as the short-term level of speech fluctuates.

The overall error rate was quite similar between the AMS-based classifier and the tested hearing aids. However, while classification with AMS patterns is based on relatively short segments of the input signal (500 ms), the classification decisions of the hearing aids were measured based on average hearing aid output over 30 seconds. It is important to note that the classification evidence can be accumulated across the time for achieving higher accuracy, but this has to be paid for with more sluggish program adaptation.

When tested with similar (but different) material as in the training set (9 hours in total, no speech pauses, 1970s pop music, reverberant speech included), an overall recognition rate of 92.1% for 500 ms segments was achieved with the AMS-based approach.

Even if the training of neural network used for pattern recognition could be computed outside a hearing aid, the proposed classification approach is still way too expensive in terms of computational effort and memory to be implemented in present hearing aids. In contrast to other potential applications of such a sound classification systems, considerable effort would be necessary to simplify feature extraction and to reduce the complexity of the neural network without loosing performance.

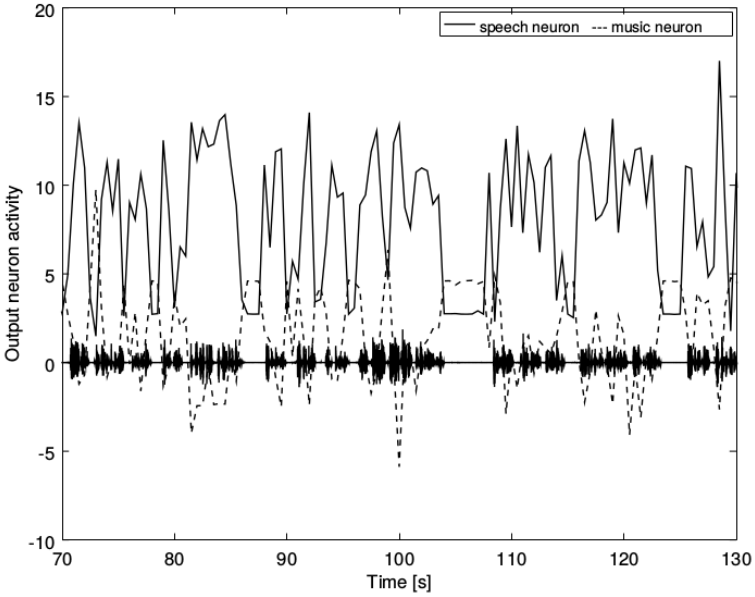

Figure 1 – Time signal of a 60 s segment of the speech input signal (poem), together with the corresponding output neuron activities for speech and music, respectively. The output neuron activities for the other three classes were omitted for clarity.

Acknowledgements

The authors thank Gibak Kim [11] for sharing the Matlab code for AMS feature extraction.

References

[1] GJERDINGEN, R. O., , and D. PERROTT: *Scanning the dial: The rapid recognition of music genres*. *Journal of New Music Research*, 37(2), pp. 93 – 100, 2008.

[2] KATES, J. M.: *Classification of background noises for hearing aid applications*. *Journal of the Acoustical Society of America*, 97(1), pp. 461–470, 1995.

[3] HAMACHER, V., J. CHALUPPER, J. EGGERS, E. FISCHER, U. KORNAGEL, H. PUDER, and U. RASS: *Signal processing in high-end hearing aids: State of the art, challenges, and future trends*. *EURASIP Journal on Applied Signal Processing*, 18, p. 2915–2929, 2005.

[4] BÜCHLER, M., S. ALLEGRO, S. LAUNER, and N. DILLIER: *Sound classification in hearing aids inspired by auditory scene analysis*. *EURASIP Journal on Applied Signal Processing*, 18, p. 2991–3002, 2005.

[5] PITA, R. G., D. AYLLON, and J. RANILLA: *A computationally efficient sound environment classifier for hearing aids*. *IEEE Transactions on Biomedical Engineering*, 62(10), pp. 2358 – 2368, 2015.

[6] BAUMANN, S., O. JOLY, A. REES, C. I. PETKOV, L. SUN, A. THIELE, and T. D. GRIFFITHS: *The topography of frequency and time representation in primate auditory cortices*. *eLife*, 4, 2015. doi:10.7554/eLife.03256.

[7] TCHORZ, J. and B. KOLLMEIER: *SNR estimation based on amplitude modulation analysis with applications to noise suppression. IEEE Transactions on Speech and Audio Processing*, 11(3), pp. 184 – 192, 2003.

[8] HEALY, E. W., S. E. YOHO, J. CHEN, Y. WANG, and D. WANG: *An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type. J. Acoust. Soc. Am*, 138(3), pp. 1660 – 1669, 2003.

[9] PITT, M., L. DILLEY, K. JOHNSON, S. KIESLING, W. RAYMOND, E. HUME, and E. FOSLER-LUSSIER: *Buckeye Corpus of Conversational Speech (2nd release) [www.buckeyecorpus.osu.edu]*. Department of Psychology, Ohio State University, Columbus, OH, 2007.

[10] HOLUBE, I., S. FREDELAKE, M. VLAMING, and B. KOLLMEIER: *Development and analysis of an international speech test signal (ISTS). International Journal of Audiology*, 49(12), pp. 891 – 903, 2010.

[11] KIM, G., Y. LU, Y. HU, and P. C. LOIZOU: *An algorithm that improves speech intelligibility in noise for normal-hearing listeners. Journal of the Acoustical Society of America*, 126(3), pp. 1486 – 1494, 2009.