Modulation Feature Extraction for Robust Automatic Speech Recognition

Niko Moritz[1], Jörn Anemüller[2], Birger Kollmeier[1,2]

[1] *Fraunhofer IDMT, Project Group Hearing, Speech and Audio Technology, 26129 Oldenburg,*
E-Mail: niko.moritz@idmt.fraunhofer.de

[2] *Medical Physics, Dept. of Physics, Carl-von-Ossietzky University Oldenburg, 26129 Oldenburg*

Introduction

Today's most commonly used features for automatic speech recognition (ASR), such as the Mel-Frequency Cepstral Coefficients (MFCCs) and Perceptual Linear Predictive analysis (PLPs), use the spectral envelope as the prime carrier of the phonetic identity for the classification process. However, the spectral envelope can easily be disrupted by additive and convolutional noise, whereas human speech perception is much less susceptible by distortions.

Perceptual experiments indicate that the human auditory system analyzes modulation components of a received speech signal [1], a process that can be mimicked in signal processing by analyzing longer time trajectories of the spectral envelope [2]. The temporal dynamics of a speech signal, i.e. amplitude modulations, seem to serve more reliable cues for robust speech recognition [3-6].

In this contribution we provide studies in modulation frequency analysis for an environmental noise robust feature extraction method. Several variants of modulation extraction schemes have been investigated using the Aurora-2 benchmark task and optimization for parameters such as length of temporal analysis window has been performed.

Modulation Feature Extraction

In this contribution modulation feature extraction is based on the amplitude modulation spectrogram (AMS) [7]. The audio signal is spectrally analyzed by a short-time Fourier transform (STFT) and by squaring the magnitude of the complex spectrogram values the power spectral density is obtained. Then certain frequency regions are decomposed into a set of critical bands according to the bark frequency scale. In the next step longer time trajectories within each band of the spectral envelope are analyzed by a second STFT. Thus, a representation of the amplitude modulations for each center frequency band is obtained. Based on perceptual [8-10] and ASR experiments [3,4] modulation frequency components outside the range between 2 and 16 Hz are discarded. This restriction already allows the sorting out of influences that were not caused by the speech signal itself [5]. In the last step the real or imaginary part of the remaining complex AMS coefficients is taken, which is then normalized to the unit circle and multiplied by the compressed length of the complex pointer. By the use of the real or imaginary part phase information are preserved in the resulting AMS coefficients. For the compression of the complex pointer a third root function is suggested, since a logarithmic compression for instance can cause negative values, which would lead to unwanted interactions with negative values of the real or imaginary part.

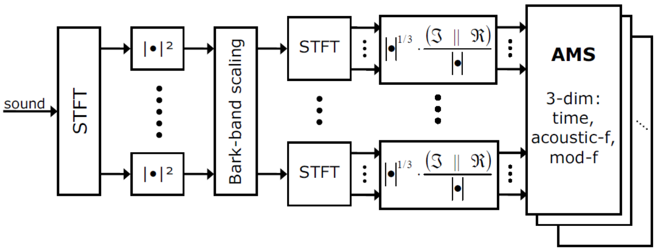
sound
STFT
|•|²
|•|²
Bark-band scaling
STFT
STFT
$|\bullet|^{1/3} \cdot \frac{(\Im \,\|\, \Re)}{|\bullet|}$
$|\bullet|^{1/3} \cdot \frac{(\Im \,\|\, \Re)}{|\bullet|}$
AMS
3-dim: time, acoustic-f, mod-f

Figure 1: Signal processing scheme for computing AMS features.

AMS Parameter Settings

Default AMS parameters used within this work are: i) The analysis window of the first STFT has length 25 ms and shift 10 ms; ii) The window of the second STFT has length 310 ms, which emerged from experiments as the optimal window length [6]. The shift remains fixed to 10 ms, implying oversampling; iii) The length of the feature vectors is reduced to 39 by a principal component analysis (PCA) to ensure the comparability to MFCCs that are used as baseline. In the further reading these parameter settings are abbreviated with AMS_{310}.

Experimental Setup

The experiments with the presented AMS features are performed using the Aurora-2 framework [11]. The Aurora-2 data is based on TIDigits (clean samples of spoken English digit strings) downsampled to 8 kHz. Eight different noise types were added to the clean speech data at signal-to-noise ratios (SNRs) ranging from 20 dB to -5 dB in 5 dB steps. The data were split into two training sets and three test sets. The training sets differ in a clean- and a multi-condition training. The test sets A and B each comprise four different types of noise. The division between test set A and B is related to the multi-condition training, since the noise types used for set A are also used to add noise to the clean training data to create the multi-condition training set. For test set C speech data with one noise type from test set A and one from B were convolved with a filter that simulates the behavior of a telecommunication terminal. The isolated word recognition engine of the Aurora-2 framework is based on linear HMMs using 18 states per word (including the two non-emitting states) and mixtures of three Gaussians per state. Baseline results are obtained with MFCCs, which comprise 12 cepstral coefficients (without the 0th coefficient) and the logarithmic frame energy plus the corresponding delta and acceleration coefficients (resulting in 39 features for each frame).

Recognition results shown within this contribution only refer to the clean-condition training.

Table 1: Average WERs for the different Aurora-2 test conditions and for different AMS feature versions. $AMS_{310.r3}$ represents the AMS version of Figure 1, in which the real or imaginary part of the AMS is neglected. $AMS_{310.R}$ and $AMS_{310.I}$ represent the AMS versions once with the real part and once with the imaginary part. Baseline results are given by MFCCs. MFCCs + $AMS_{310.I}$ is the concatenation of both feature types. R.I. gives the relative improvement of MFCCs + $AMS_{310.I}$ compared to baseline (MFCC).

ØWER		$AMS_{310.r3}$	$AMS_{310.R}$	$AMS_{310.I}$	MFCCs + $AMS_{310.I}$	R.I.
For set A + B	Clean	11,10	5,55	5,87	0,98	6,92
	20 dB	12,05	6,69	6,46	1,74	63,81
	15 dB	15,26	8,46	8,84	4,03	69,04
	10 dB	25,88	19,44	18,54	11,54	66,97
	5 dB	46,95	41,67	33,45	30,08	52,74
	0 dB	70,57	72,90	60,84	65,68	22,07
	-5 dB	86,30	92,73	90,84	89,22	2,91
0-20 dB	Set A	34,83	30,07	25,26	22,36	50,18
	Set B	33,45	29,60	25,99	22,87	59,66
	Set C	51,45	38,01	28,93	20,55	47,65

Emphasizing Phase Information of Amplitude Modulations

This section demonstrates that phase information of modulation frequencies contain important cues for speech recognition (q.v. [4]). Differences in word error rates (WER) between different versions of the AMS features are presented and discussed. Therefore, three different AMS versions are used, as seen in Table 1. In $AMS_{310.r3}$ the complex AMS coefficients are compressed by a third root but not multiplied by the normalized real or imaginary part, whereby phase information are dropped. By comparing these results with the $AMS_{310.R}$ and $AMS_{310.I}$ it can be seen that the additional phase information within the real and imaginary part significantly improves recognition results. Furthermore, it can be observed that the imaginary part offers advantages over the real part. This property can be explained by considering the modulation transfer functions (MTF) of the used Fourier basis functions in Figure 2. The transfer function of the real part of the Fourier basis function with the center frequency at 3,125 Hz illustrates, that this part has non-zero mean and, thus, does not attenuate frequencies below this center frequency in contrast to the imaginary part (q.v. [6] for a more detailed discussion).

By concatenating $AMS_{310.I}$ features with MFCCs further significant improvements in WERs are achieved, which can be seen in Table 1. This feature combination reached a total relative improvement of 53,5 % for clean-condition training.

Conclusions

The results demonstrate that preserving phase information in modulation features can significantly increase ASR performance. In this context it is shown that an odd analysis basis function, which has zero mean, has advantages for modulation frequency analysis. This observation can be explained by analyzing the corresponding MTFs.

It is useful to combine the dynamic long-term features of the AMS, with short-term MFCC information. Hence, short-term features, that generally achieve very good ASR performances for clean speech, benefit from significantly increased noise robustness, due to the additional long-term information. On the Aurora-2 task, the best feature set presented in this contribution reached an overall relative improvement of 53,5 % on clean-condition training.

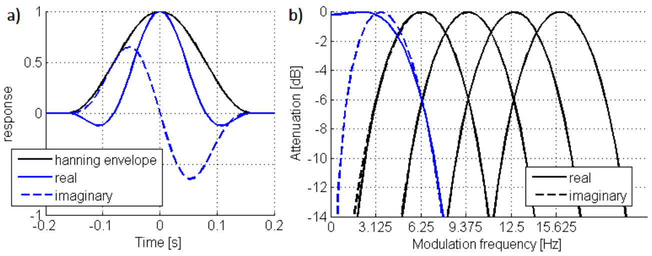

Figure 2: a) Fourier basis function with 3,125 Hz center frequency. b) Normalized frequency response of the Fourier basis functions used to obtain the AMS_{310}.

References

[1] T. Dau, and B. Kollmeier, “Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers,” *J. Acoust. Soc. Am.* 102(5), pp. 2892-2905, 1997.

[2] B. Kollmeier, and R. Koch, “Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction,” *J. Acoust. Soc. Am.* 95(3), pp. 1593-1602, 1994.

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

[4] N. Kanedera, H. Hermansky, and T. Arai, “On properties of modulation spectrum for robust automatic speech recognition,” *Proc. ICASSP* 1998, pp. 613-616, 1998.

[5] H. Hermansky, and N. Morgan, “RASTA processing of speech,” *IEEE Trans. on Speech and Audio Processing* 2(4), pp. 578-589, 1994.

[6] N. Moritz, J. Anemüller, and B. Kollmeier, „Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments,“ *Proc. ICASSP* 2011, 2011.

[7] B. Kollmeier, and R. Koch, “Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction,” *J. Acoust. Soc. Am.* 95(3), pp. 1593-1602, 1994.

[8] R. Drullman, J.M. Festen, and R. Plomp, “Effect of temporal envelope smearing on speech reception,“ *J. Acoust Soc. Am.* 95, pp. 1053-1064, 1994a.

[9] R. Drullman, J.M. Festen, and R. Plomp, “Effect of reducing slow temporal modulations on speech reception,“ *J. Acoust Soc. Am.* 95, pp. 2670-2680, 1994b.

[10] T. Arai, M. Pave, H. Hermansky, and C. Avendano, “Intelligibility of speech with filtered time trajectories of spectral envelopes,” *Proc. ICSLP* 96, 1996.

[11] H.G. Hirsch, and D. Pearce, “The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions,” In: *ISCA ITRW ASR*, 2000.