ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Amplitude Modulation Filters as Feature Sets for Robust ASR: Constant Absolute or Relative Bandwidth?

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

[1]Fraunhofer IDMT, Project group Hearing, Speech and Audio Technology, Oldenburg, NI, Germany
[2]University of Oldenburg, Medical Physics Department, Oldenburg, NI, Germany
niko.moritz@idmt.fraunhofer.de

10.21437/Interspeech.2012-379

Abstract

Many research efforts in the field of feature extraction for automatic speech recognition are focused on analyzing slow amplitude fluctuations of speech. In this study the importance of spectral and temporal resolution for the amplitude modulation frequency analysis are investigated in order to provide guidance for the appropriate filter design. Therefore, different wavelet and Fourier transform like filter time scales are examined, i.e. the importance of time and frequency separation is compared. The results demonstrate that analyzing three separate amplitude modulation frequency bands of constant absolute bandwidth that cover the range from about 2 to 16 Hz are sufficient for automatic speech recognition.

Index Terms: amplitude modulation, speech recognition, wavelet transform, feature extraction

1. Introduction

Recently, electronic devices that are using automatic speech recognition (ASR) increasingly enter into everyday life. This can be observed in particular for mobile phones and tablet computers that nowadays almost all have an application for voice control. Moreover, game consoles and TVs start using ASR with microphones situated far away from the speaker. Since far-talk microphones catch up not only the desired signal but also noise and reverberation from the room, it is obvious that these systems have high demands on ASR robustness. Although these systems already show good performance not only under laboratory conditions but also in practice, there is still much room for improvements. This becomes obvious as soon as such systems are compared to human speech recognition (HSR) capabilities.

An area in ASR that still promises room for increasing noise and distortion robustness can be found in feature extraction. In [1], for example, a gap between HSR and ASR performance of 15 dB in terms of the signal-to-noise ratio (SNR) is determined. In this study the major discrepancy of 10 dB SNR is traced back to the ASR front-end, i.e. the feature extraction and signal preprocessing.

A common approach to improve ASR feature extraction methods is to mimic signal processing strategies of the human auditory system. The most commonly used acoustic features for ASR are the Mel-frequency cepstral coefficients (MFCCs) that, however, use relatively little knowledge of the auditory system (e.g. Mel-frequency warping and log compression). An important difference of MFCCs to human signal processing is that MFCCs do not model higher order processing stages of the auditory pathway like analyzing slow amplitude fluctuations. Since it is known that amplitude modulations (AM) carry essential information for speech perception [2, 3], it seems to be important to realize such processing steps in the feature extraction. Delta and acceleration coefficients that are usually used in combination with MFCC features are a first attempt to analyze a larger temporal domain of approximately up to 100 ms [4]. However, compared to the auditory filters found in the human auditory system, this is a very limited domain [5-7].

For this reason several new feature extraction methods have been investigated in the past that try to emphasize the important modulation frequency domain of speech by applying different types of band-pass filters. The early methods developed used just one modulation filter. Two well-known examples are the RelAtive SpecTrAl (RASTA) processing [8], which is a filter that emphasizes components between about 1 and 12 Hz and the modulation spectrogram by Kingsbury *et al.* [9]. In more recent work, feature extraction methods tend to use separable band-pass filters spread along the important modulation frequency domain of speech. In [10], for example, a discrete cosine transform is used to analyze different AM frequencies of critical bands that were previously smoothed by frequency-domain linear prediction. Furthermore, two dimensional Gabor filters can be used for feature extraction to analyze different spectral, temporal and spectro-temporal modulation frequency components of a spectrogram [11, 12].

In order to analyze the important AM frequency domain of speech that ranges from approximately 2 to 16 Hz [5-7], temporal filter extensions of approximately 60 to 500 ms are needed. This large variance in necessary filter time scales raises the question, whether a filter design with a frequency resolution matched to its analyzed center frequency (CF) is mandatory. Therefore, in this contribution the wavelet transform (WT) with Morlet basis functions is used for the AM analysis and different modifications of the filter bandwidths are examined. The basic feature extraction method that is used for this study is the amplitude modulation spectrogram (AMS) [13].

2. Amplitude modulation spectrogram

AMS are motivated by psycho-physical and psycho-physiological findings on the processing of amplitude modulations in the auditory system of mammals. Langner *et al.* suggested the existence of a periodotopic arrangement of neurons in the inferior colliculus (IC) that are tuned to certain modulation frequencies [14]. These neurons were found to be almost orthogonal to the tonotopic arrangement of neurons that are tuned to certain center frequencies. Psychoacoustic studies of Dau *et al.* account for the theory of the modulation frequency analysis of each center frequency band [15]. In [16] these findings are introduced to signal processing by definition of the AMS, which form the basis for the features used in this study. However, it should be

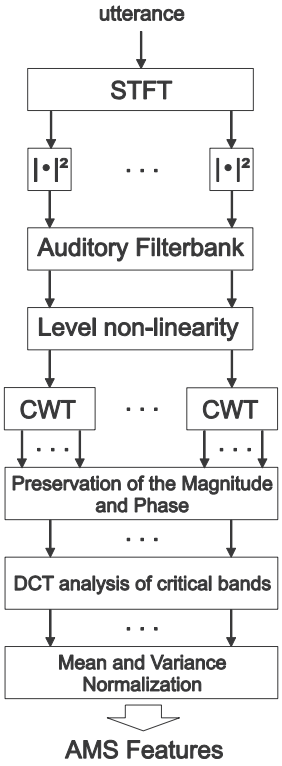

Figure 1: *Block diagram of the signal processing steps for computing AMS features.*

noted that the original AMS of Kollmeier and Koch are used here in a modified version. The block diagram in Figure 1 depicts the signal processing steps to compute the new AMS.

In the first stage a spectrogram is computed by using the short-time Fourier transform (STFT) with 25 ms blocklength and 10 ms shift. Squaring the magnitude values of the complex spectrogram provides the spectral Hilbert envelope. The auditory filter bank in the next stage performs a decomposition of acoustic frequencies into critical bands according to the Mel-frequency scale. A level compression is conducted on the Mel-spectrogram, which mimics the compression of the outer hair cells in the inner ear. For this, we apply the non-linearity function

$$f_C(x) = \left(x^{0.4} + \log(x) + 1\right)/2 \qquad (1)$$

where x denotes the values after the Mel-band decomposition. The function (1) has demonstrated to increase ASR robustness compared to more well-known compressing functions like the logarithm. However, one disadvantage compared to the logarithm is that multiplicative components such as level scaling of the input signal do not become additive after the application, due to the exponential component $x^{0.4}$ in function (1). Therefore, energy normalization is required to compensate for this disadvantage. A way to circumvent this problem is to scale each utterance to +/- 1.

The next step in Figure 1 involves the analysis of temporal amplitude fluctuations by a WT for each critical frequency band. Therefore, complex Morlet basis functions are used that are tuned to different modulation frequencies that approximately cover the domain from 2 to 16 Hz (see Figure 2).

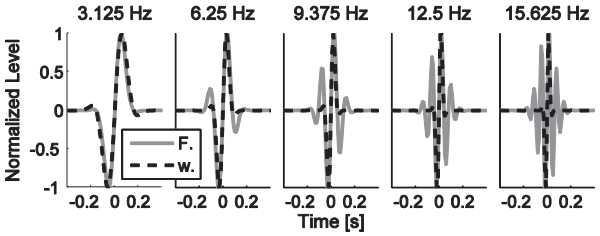

Figure 2: *Imaginary part of complex Morlet wavelet functions for different center frequencies. The solid gray lines depict the filter functions with a constant absolute bandwidth, whereby the dashed lines depict the corresponding wavelet like approach.*

Because of relatively long filter time scales for the modulation frequency analysis of speech (up to several hundred milliseconds) it is necessary to preserve the phase and thus location information about the analyzed frequency components within each filtered block. This can be achieved by taking the imaginary or real part of the filter output, whereby magnitude and phase information remain coded in the coefficients. Here we take the imaginary part, since its mean is zero.

For the decorrelation and a following reduction of the AMS coefficients a discrete cosine transform (DCT) is used in a related way as it is known from MFCC computation. For this the spectral vector of each modulation frequency band is separately analyzed by the DCT. The DCT coefficients of each AM band are then reduced to ten. Thus, the number of features for each frame is equal to ten times the amount of analyzed AM frequency bands. In the last step a mean and variance normalization (MVN) is performed separately for each utterance.

3. The experimental framework

The test environment for the proposed AMS features in this study is the Aurora 2 framework [17] that contains both the speech data as well as the specifications for the hidden Markov model (HMM) classifier. The speech data is based on TIDigits, where English connected digits are spoken by native male and female speakers of different age. Aurora 2 provides a clean- and multi-condition training set. Since TIDigits are originally clean recordings sampled at 20 kHz, for Aurora 2 the data is resampled at 8 kHz and mixed with eight different Noisex-92 noise types at different signal-to-noise ratios (SNR). The noise types are subway, babble, car, exhibition (that are used for test set A and also during multi-condition training), restaurant, street, airport and train-station (that are used in test set B). The SNR conditions range from 20 dB to -5 dB in 5 dB steps. In addition to test set A and B a test set C exists, where the clean and noisy test data of subway and street noise are convolved with an impulse response that simulates the behavior of a telecommunication terminal.

The Aurora 2 back-end consists of whole-word HMMs with 16 emitting states per word and three Gaussian mixture components per state. The HMMs are left-to-right models without skips over states. In addition to the word HMMs two silence models are used. One is for the start and end silence of each utterance and one is for sort-pauses between words. The training and testing is performed using the Hidden Markov Model Toolkit (HTK) version 3.1.1.

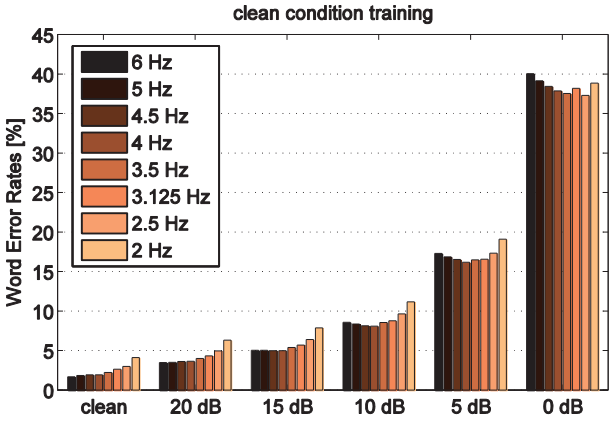

Figure 3: A*verage WERs for different SNR conditions and different constant absolute -3 dB BW settings using AMS features. The AM frequency filters that are used sample the modulation frequency domain range from 3.125 Hz to 15.625 Hz in 3.125 Hz steps.*

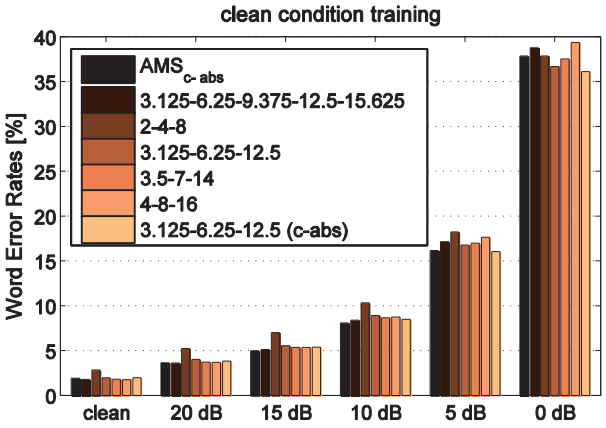

Figure 4: A*verage WERs of AMS features using AM filters with different BW settings. For the wavelet like filter approach the filter CFs are shown separated by minus signs. An exception is denoted by (c-abs) and suggest that here a constant absolute BW of 4 Hz is used.*

4. Experiments

The presented experiments focus on the difference between band-pass filters with constant absolute and constant relative time scales according to the analyzed AM frequency component. As described in Section 2, we use the imaginary part of the complex Morlet basis function for the analysis. This is a sine function with a certain center frequency (CF), which is modulated by a Gaussian window (see Figure 2). The filters' CFs for the experiments with constant absolute BWs are ranging from 3.125 Hz to 15.625 Hz in 3.125 Hz steps. These are the same CFs as for the 32-point fast Fourier transform (FFT) that is used for the STFT based AMS version in [13].

Figure 2 exemplary represents the shape of the basis functions for a constant absolute -3 dB bandwidth (BW) of 4 Hz and for the corresponding wavelet like filter design (i.e. a constant-Q filter approach). It can be seen that the temporal localization of the constant-Q filter approach increases with higher CFs, whereby its frequency resolution decreases because the filter BW is inversely proportional to the time scale.

Since relatively broad temporal filter extensions are needed to analyze the slow amplitude fluctuations of speech (cf. Figure 2), 15 frames of the feature stream at the beginning and end of each utterance are truncated to reduce the influence of bad artifacts that may occur due to an improper boundary criterion (which here is zero padding). The number of frames that are truncated is equivalent to a 300 ms analysis window as it is used for the AMS version in [13].

4.1. Constant absolute filter bandwidth

Figure 3 represents the ASR results for AMS that use a constant absolute BW for each AM filter. To estimate the optimal average -3 dB BW, the ASR results for different BW settings are tested and presented. On the one hand, the results demonstrate that increasing the BW, i.e. increasing temporal localization, leads to reduced WERs for clean speech but also reduce noise robustness. On the other hand, increasing the resolution in AM frequency space seems to increase noise robustness. However, the counteraction of these two properties results in a minimum WER at about 4 Hz BW for SNRs less than 15 dB. Therefore, 4 Hz is assumed to be the optimal average -3 dB BW and in the further reading the according feature set is abbreviated by $AMS_{c\text{-}abs}$.

4.2. Constant relative filter bandwidth

In this section AM frequency filters of constant relative BW are used, which corresponds to the idea of the wavelet transform and constitutes the main difference to the Fourier transform. Figure 4 represents the ASR results for this experiment. The comparison of WERs for the AMS with constant absolute and relative filter time scales, which both sample the AM frequency domain ranging from 3.125 Hz to 15.625 Hz in equal distant steps, demonstrates that the WT based approach has no advantage compared to the Fourier transform based approach. Here, even the opposite is the case, which means that the noise robustness is decreased for the constant-Q filter approach, while the ASR performance for clean speech is slightly increased. However, for uniformly distributed AM frequency filters, the constant-Q filter approach effects a strong oversampling in modulation frequency space, which might be a disadvantage due to redundant information in different filter outputs. This fact is illustrated in Figure 5. To exploit the full possible advantage of the WT, therefore, the AM filters need to be distributed in AM frequency domain in such a way that a greater overlap of each filter is avoided as it is known from the dyadic wavelet transform, for example. Furthermore, the filter distribution should still concentrate on the important AM frequencies of speech (i.e. approx. 2-16 Hz) to suppress non-speech components in the signal. Figure 5 depicts an example for such a filter distribution in modulation frequency domain and in Figure 4 the corresponding ASR results are illustrated. The results indicate that this filter distribution cannot outperform the $AMS_{c\text{-}abs}$ approach but achieves comparable results (see results for 3.125-6.25-12.5), while the number of features is reduced due to a less number of AM frequency filters. Thus, the number of analyzed AM frequency bands is reduced from five to three, whereby the total feature number is cut down from 50 to 30.

Figure 4 also illustrates the results for the $AMS_{c\text{-}abs}$ without using the 9.375 and 15.625 Hz filters to determine whether only the

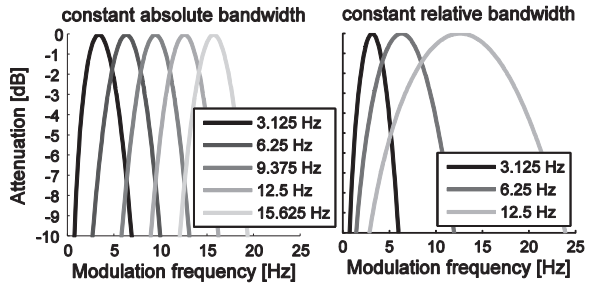

Figure 5: *Normalized frequency response of AM filters with a constant absolute -3 dB BW of 4 Hz (left panel) and for an appropriate constant-Q filter approach (right panel).*

constant-Q filter approach allows the number of AM bands to reduce. The comparison of both variants shows that also the constant absolute BW approach achieves similar or even slightly better WERs for a reduced number of AM frequency filters. Further attempts to cut down the number of AM bands led to significant worse WERs, without showing the results here in detail due to space limitations. This implies that three AM frequency filters of approximately 4 Hz BW each, which cover the AM frequency domain from 2 to 16 Hz, provide sufficient time and frequency separation for AM frequency feature extraction.

4.3. Comparison to the state-of-the-art

In this section the long-term spectral information of AMS are combined with short-term spectral information of MFCCs by concatenation and the recognition scores are compared to other state-of-the-art feature extraction methods. Table 1 represents the results. It is demonstrated that the tested AMS feature set achieved a total relative improvement in WERs of 49.61 % over the Aurora 2 baseline.

5. Conclusions

The presented study provides indications for the appropriate design of amplitude modulation frequency filters for ASR feature extraction methods. It is demonstrated that a filter configuration similar to auditory filters does not improve speech recognition accuracies compared to a configuration that uses a constant absolute bandwidth for each filter. However, in order to achieve good ASR noise robustness, it is important to be limited only to the AM frequency range from 2 to 16 Hz. The results indicate that it is sufficient to divide this frequency range into three separate frequency bands with a -3 dB bandwidth of about 4 Hz each. On the Aurora 2 task the presented AMS features achieved a total relative improvement of 49.61 %.

6. References

[1] B. T. Meyer, T. Brand, and B. Kollmeier, "Effects of speech-intrinsic variations on human and automatic speech recognition of spoken phonemes," *J. Acoustic. Soc. Am.* 129, pp. 388-403, 2011.

[2] T. Houtgast, and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acoustica* 28, pp. 66-73, 1973.

[3] T. Houtgast, and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoustic Soc. Am.* 77, pp. 1069-1077, 1985.

Table 1. *The average WER is determined over all test conditions for SNRs ranging from 0-20 dB. Relative improvement gives the total relative improvement over baseline (MFCC). MFCC denotes MFCCs using the energy coefficient plus delta and acceleration coefficients. CMN means cepstral mean normalization, which is performed on MFCCs including the delta and acceleration coefficients (note that this is different to HTK and leads to better WERs). $AMS_{3c\text{-}abs}$ depicts the AMS version using three AM filters of constant 4 Hz BW sampling the AM frequencies 3.125, 6.25, and 12.5 Hz.*

	Average WER (clean training)	*Average WER (multi training)*	*Relative Improvement*
MFCC	40.87 %	14.49 %	–
$MFCC_{+CMN}$	28.28 %	12.01 %	25.99 %
$AMS_{3c\text{-}abs}$ + $MFCC_{+CMN}$	14.37 %	8.68 %	49.61 %

[4] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust. Speech Signal Process.* 34(1), pp. 52-59, 1986.

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

[6] 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, 1994.

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

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

[9] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication* 25, pp. 117-132, 1998.

[10] S. Ganapathy, S. Thomas, and H. Hermansky, "Comparison of modulation features for phoneme recognition," *Proc. ICASSP* 2010, pp. 5038-5041, 2010.

[11] M. Kleinschmidt, "Localized spectro-temporal features for automatic speech recognition," In: *Proc. Eurospeech*, 2003.

[12] B. T. Meyer, S. V. Ravuri, M. R. Schädler, and N. Morgan, "Comparing different flavors of spectro-temporal features for ASR," *Interspeech*, pp. 1269-1272, 2011.

[13] N. Moritz, J. Anemüller, and B. Kollmeier, „Amplitude modulation spectrogram based features for robust speech recognition in noisy and reverberant environments," *Proc ICASSP* 2011, pp. 5492-5495, 2011.

[14] G. Langner, and C.E. Schreiner, "Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms," *J. of Neurophysiology* 60, pp. 1799-1822, 1988.

[15] 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.

[16] 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.

[17] 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.