ZERO-CROSSING-BASED TEMPORAL SEGMENTATION AND CLASSIFICATION OF AUDIO SIGNALS

Konstantin Biatov, Martha Larson, Stefan Eickeler

Fraunhofer Institute for Media Communication (IMK), Sankt Augustin, Germany
{biatov, larson, stefan.eickeler}@imk.fraunhofer.de

ABSTRACT

The task of audio classification is streamlined if the audio stream can be decomposed in real time into homogeneous segments larger than the individual frame. If segment morphology reflects the underlying audio type, classifier design and training becomes more transparent and can directly exploit real-world knowledge. The experiments presented here explore real-time zero-crossing-based audio stream segmentation and classification. We segment an un-edited audio stream from a German-language radio program into three classes: speech, speaker breathing and music. Features related to modulation of the amplitude envelop are exploited to segment the audio stream into homogeneous sub-syllable segments. The segments prove useful for classifying the audio-stream, especially the speech portions. We use features extracted from a zero-crossing-based pseudo-spectrum to train a Gaussian classifier that classifies audio content into the three categories. Our results demonstrate that our zero-crossing-based segmentation and classification method is a viable one, yielding satisfactory performance in real time.

1 INTRODUCTION

Progress of audio processing technologies has led to an expanding role for audio as an information source and a communication medium. The increasing importance of audio data has inspired the vision that that audio should be just as easily segmentable, indexable, searchable and even as translatable as text. Motivated by this vision is the move away from frequency-domain and other calculation intensive approaches, wherever there are not justified by the task or the medium.

This paper is an exploratory investigation in the direction of real-time segmentation and classification of audio data. This work was driven by two central insights. First, audio processing does not have to take place on the frame level, but rather that benefit is to be derived by matching the method to the medium. In our experiments we use modulations in the amplitude envelop to first decompose the audio stream into segments motivated by signal morphology, which in the case of speech turn out to be sub-syllables. Second, audio processing does not have to take place in the frequency domain. We introduce a pseudo-spectrum derived from the distribution of zero-crossings and show that features can be extracted from it that yield reliable classification into audio categories. We segment an unedited audio stream from a German-language radio program into three classes: speech, speaker breathing and music.

Our experiments draw motivation from the literature on segmentation and classification based on larger-than-frame segments. A two pass classification was introduced in [1]. In the first pass, speech/non-speech boundaries were detected using a reduced phoneme inventory and gender-independent acoustic models. The second pass identified speaker changes, using only the boundaries detected in the first pass as potential candidates. Although this approach increased segmentation accuracy, it also increased segmentation time.

Several approaches have decomposed speech into syllable and syllable-like units. An automatic segmentation of speech into syllable units based on the convex hull of the loudness function in described in [2]. Relative loudness maximums are interpreted as potential syllabic peaks and relative loudness minimums as potential syllabic boundaries. A local loudness minimum separated from another local minimum by less than *100 ms may be insignificant, but a minimum of the same* magnitude which had no other minimums within 500 ms would indicate a syllable boundary.

Automatic detection of syllable boundaries based on full-band energy contour and voicing detection using modified autocorrelation is described in [3]. The syllable boundary is determined to be the point of minimum energy contour within each portion of consecutive possibly unvoiced frames.

[4] describes a new representational format for speech, the modulation spectrogram, that represents amplitude modulation frequencies in the speech signal between 0 and 8 Hz, with the peak at 4 Hz corresponding to modulation spectrum of speech. The modulation spectrogram robustly extracts information related with the syllabic segmentation of speech.

The literature on zero-crossing rate is a growing corpus, and also served as an inspiration for our approach. Possibilities for spectral analysis based on zero-crossing are described in [5]. In [6] a spectral analyzer based on zero-crossing is demonstrated. Using zero-crossing interval measurement for formant frequency estimation in noise is presented in [7]. Experiments on speech recognition in a noisy, real–world environment in which the frequency information of the signal is obtained from zero-crossing intervals are described [8]. In [9] experiments which use gravity centers of energy as additional feature to the classical set of MFCC in automatic speech recognition system are described. These experiments demonstrate improved recognition performance when gravity centers are computed from zero-crossing intervals detected at the output of the filters of an ear model.

Using just zero-crossings for speech/music discrimination has been investigated previously in [10]. Other research on speech/music discrimination and classification has used temporal information from speech waveform and as well as spectral information. Spectral information has include such features as 4 Hz modulation energy, rolloff of the spectrum, spectral flux and spectral centroid. In this paper we would like to demonstrate that features like these can be extracted from a zero-crossing-based pseudo-spectrum and used to classify and audio stream with satisfactory accuracy.

In the next section of this paper, section 2, we describe our zero-crossing-based downsampling of the time-signal and present motivation for the approach. We introduce our method for extracting sub-syllable segments by looking at modulations in the amplitude envelop. We detail the calculation of the zero-crossing-based pseudo-spectrum and the features we derive from it. In section 4 we describe our classification experiments. Section 5 presents conclusions and outlook.

2 PSEUDO-SPECTRAL ANALYSIS

2.1 Zero-crossing-based Smoothing

Assuming that signal is normalized, i.e. the mean has been removed, we define zero crossings, t_n, $n = 1,2,3...$, as the times at which the signal changes sign or is equal to zero. The successive zero-crossing intervals are defined as $z_n = t_n - t_{n-1}$. As discussed in [7] the successive zero-crossing intervals of sinusoidal signal exhibit high consistency and are inversely related to frequency. The signal resulting from the sum of more than one sinusoidal component, however, may not satisfy this principle. Usually the speech signal between two zero-crossings has more than one sinusoid. To be sure that the representation of speech signal as one sinusoid between two zero-crossings is valid for the purpose of discrimination we have implemented the method described in [12]. The audio signal is represented as a sequence of successive zero-crossing intervals and the maximal signal amplitudes for those intervals where amplitude is positive and minimal amplitude for the intervals where amplitude is negative. The signal is represented as a sequence (z_n, A_n), $n=1,2,3,...$.

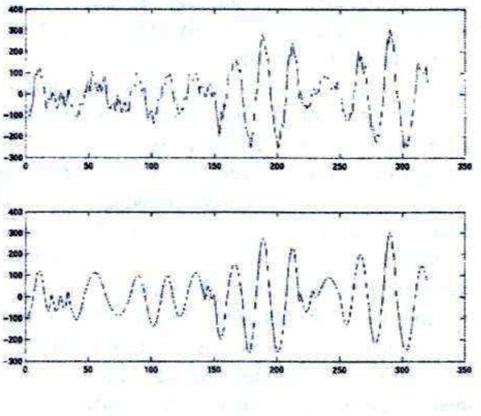

Figure 1. Comparison of original signal and recovered signals

An audio signal restored from this sequence retains sufficient quality to be understood by a human listener, and therefore holds good potential for discriminating speech, music, song, and other non-speech signals. Inconsistencies in the relationship between frequency and zero-crossing rate in the speech signal, remain well contained. From Figure 1 an impression can be gained of exactly how close the original speech signal is to the speech signal restored after zero-crossing-based downsampling.

According to Kedem [5] the zero-crossing rate of repeatedly differentiated series converges to its least upper bound. The restored signal has the same zero-crossing rate as its derivative. Since we are assured of an least upper bound, we are justified in considering that in each successive interval we have only one sinusoid and in using the inverse of the length of the interval as a frequency of the signal in this interval.

2.2 Extracting larger-than-frame segments

The changes that are relevant for segmenting and classifying a radio program, are not those that occur on a frame to frame, but rather those which occur between some higher-level content based units. Audio content processing can be streamlined if a higher-level segment can be found, which is identifiable in the signal, but whose morphology reflects underlying audio content.

Segments which are morphologically justified are also to be homogeneous. We want to exploit facts about the audio stream such as that a strain of music is never going to separate two halves of a syllable of speech. We are interested in morphological properties of the signal that correspond to detectable sections of the audio input. For breathing this might be individual breaths, and for music, the notes. Here we concentrate on extracting a basic unit for spoken audio, namely the syllable. We find that sub-syllables, units that are useful since they never straddle syllable boundaries and can be identified reliably using the minima in the amplitude envelop of the signal.

The signal is divided into frames, each 20 ms seconds long with an overlap of 10 ms. For each frame we extract the maximal non-negative amplitude from the set of amplitudes A_n, $n=1,2,3,....$ corresponding to interval between successive zero-crossing, z_n, A_n $n=1,2,3,....$ On a new signal composed from these maximum amplitudes we find the points which are local minima. These are the boundaries of our larger-than-frame units.

2.3 Deriving the Pseudo-spectrum

The zero-crossing-based pseudo-spectrum is calculated from the restored signal described in sub-section 2.1 using the following procedure. From each 20 ms frame (overlap 10 ms) we extract the lengths of the intervals between successive zero-crossings. Then extracted periods are converted into pseudo-frequencies, by $f = 1/z$, where z is the length of the interval between successive zero-crossings. For each interval between successive zero-crossings we extract the amplitude that is maximal for the interval with positive sign of the signal, and is minimal for the interval with negative sign of signal. Then using 24 frequency bands according to the Bark scale, we group frequencies that correspond to one band together and then normalize all groups using sum of all frequencies. As the result for each 20 ms frame we have smoothed a pseudo-spectrum which can be deployed for a range of discrimination tasks.

The resulting pseudo-spectrum is well smoothed, but retains the discriminative features of the original speech signal. Figure 3 illustrates the pseudo-spectrum corresponding to the 16 kHz waveform depicted in Figure 2. These figures motivate the use of the pseudo-spectrum distinguish voiced and unvoiced parts of the signal and identify the syllabic modulation of the speech signal. Figure 4 is the pseudo-spectrogram of a waveform in which the speaker breathes between words or phrases. The duration of the breath is 100 ms. Once again, visual inspection confirms that the pseudo-spectrum contains enough information about the signal to allow a discrimination to be made.

Using the amplitudes which correspond to each zero-crossing interval we calculate a value related to energy.

$$S = \frac{abs(z_n * A_n)}{2} \qquad (1)$$

S approximates the square of the amplitude of the envelop corresponding to interval z_n. We group together all S values corresponding to the same Bark band and normalize these values by dividing by total sum for all intervals. In this way we produce a pseudo-spectrum representing the distribution of energy with respect to frequency bands.

We use this pseudo-energy spectrum to calculate an additional feature, the rolloff point. We define our rolloff point to be the frequency below which 85% of the energy in the spectrum is concentrated.

Additionally we use the centers of gravity of the Bark-spectrum frequency bands, proposed in [9].

$$CG = \frac{\sum_{i=1}^{n} A_i z_i}{\sum_{i=1}^{n} A_i} \qquad (2)$$

From the pseudo-spectrum we also calculate the frequency corresponding to the maximum zero-crossing interval in each frame. The final feature that we derive from the pseudo-spectrum is the Euclidean distance between the pseudo-spectral vectors of neighboring frames.

3 EXPERIMENTS

We experimented on data from our *Kalenderblatt* database, which contains radio programs recorded from the Deutsche Welle *Kalenderblatt* series and the accompanying transcripts. The programs are five minutes in length and each contain about 650 running words. They are liberally interspersed with music and other sound effects. Although each program is narrated by a single speaker, many other voices are present in the form of interviews and original sound footage. The radio programs were downloaded from the Internet (http://www.kalenderblatt.de) and resampled to 16 kHz. The corresponding text transcripts were also downloaded and normalized. The database is described in detail in [14]. Also in [14] is a description of the semi-automatic alignment visualization tool with which we create reference segmentations and segment labels for our experiments.

For the exploratory experiments presented here, we indiscriminately chose one 5-minute program for training and one for testing. We segmented and labeled these with the semi-automatic method down to the syllable level, also marking breathing and music. In this way we generated high-quality reference labels for training and testing. The duration of training data for the experiments presented here is 5 minutes; the duration of test data is also 5 minutes.

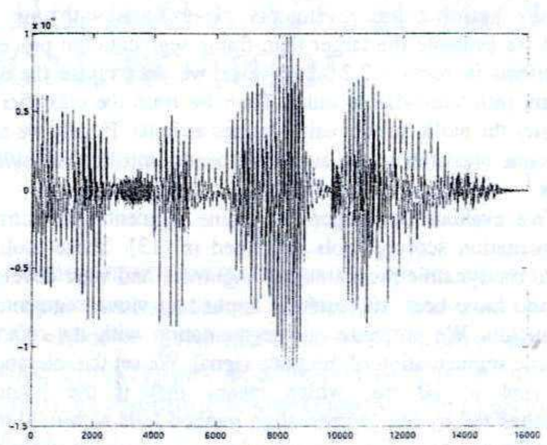

Figure 2. Waveform of speech signal. (20 ms frame with 10 ms step represented on x-axis)

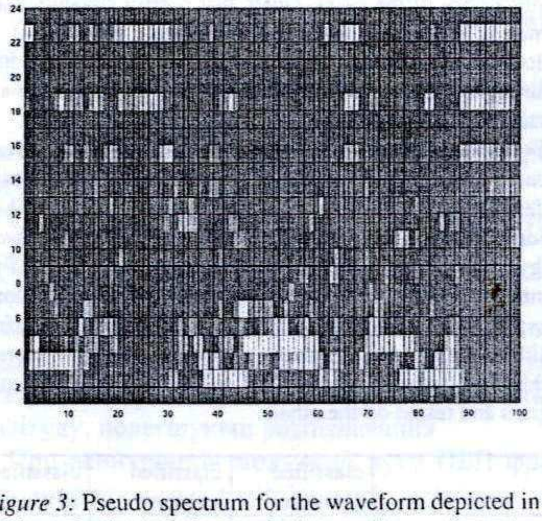

Figure 3: Pseudo spectrum for the waveform depicted in Figure 2. (Bark-scale pseudo-frequencies vs. frames)

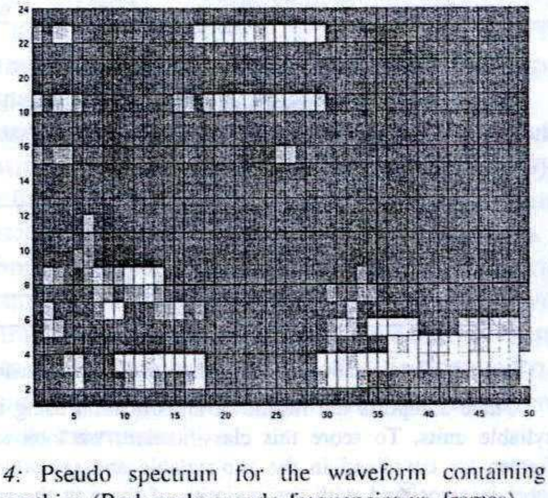

Figure 4: Pseudo spectrum for the waveform containing speaker breathing (Bark-scale pseudo-frequencies vs. frames)

We perform three evaluatory experiments with this data. First we evaluate the larger-than-frame segmentation procedure described in section 2.2 with which we decompose the audio stream into sub-syllable units. Then we train the classifier and classify the audio stream using frames as units. Finally we apply the same classifier to the audio segmented into the sub-syllable units.

We evaluate our larger-than-frame segmentation using the segmentation scoring tools described in [13]. These tools are based on dynamic programming alignment and were developed for and have been successfully applied to video segmentation evaluation. We compare our segmentation with the reference syllable segmentation of the same signal. We set the tolerance of the tool to 20 ms, which means that if the boundary hypothesized by our segmentation method falls within 20 ms of the reference boundary, the boundary is considered correct. 20 ms represents a conservatively narrow margin of tolerance, since an average syllable is 250 ms in length. Our segmentation scores 87.5% insertions and 12.5% deletions with respect to the reference segments. This means that our segments in most cases are proper sub-units of syllables and cross syllable boundaries in only 12.5% of the cases. Given that correct syllabification in German is partially a matter of convention, we feel that these results indicate that our larger-than-frame segments are indeed capturing a sub-unit that provides a useful link between signal structure and audio content.

For classification we use a maximum likelihood multivariate Gaussian classifier. The input vectors for the classifier consist of the features extracted from the pseudo-spectrum (described in section 2.3) plus their variances. These features are: average energy, gravity centers of pseudo-frequency bands, rolloff point of energy distribution, frequency corresponding to the longest interval between successive zero-crossings present in frame, Euclidean distance between pseudo-spectral vectors of neighboring frames. The classifier was trained on one 5 minute program and tested on the other.

	classified as music	classified as breath.	classified as speech
music data	**84.2%**	3.7%	11.1%
breathing data	8.2%	**89.2%**	2.6%
speech data	25.6%	12.2%	**62.2%**

Table 1: Classification on frames (20 ms, with 10 ms step)

Table 1 reports the results of classification using frames as the units. Music and breathing are classified better than speech using frame as the basic classification unit.

	classified as music	classified as speech
music data	**68%**	32%
speech data	6%	**94%**

Table 2: Classification on sub-syllables (larger-than-frame)

Table 2 reports the results of classification using the sub-syllable units. To score this classification, we look at which frames are contained in the sub-syllable and see which class label is associated with the majority of them in the reference segmentation. Speech is classified better than music using the sub-syllable units. We feel that this result reflects the utility of choosing a larger-than-frame unit that is related to the morphology of the audio content.

4 CONCLUSIONS

The zero-crossing pseudo-spectrum proposed here has proven itself to be a valuable source of features for real-time audio classification and has been shown to allow effective discrimination between speech, speaker breathing and music. Further results suggest that larger-than-frame units enhance classifier performance, but that they must be well-matched with the audio content.

Future work will focus on identifying additional features that can be derived from the pseudo-spectrum that might aid classification of audio, particularly into classes above and beyond the three we experiment with here. Finally, we hope that further work will allow us to develop a larger-than-frame segment that will be as useful for music and breath classification as the sub-syllable units are for speech classification.

5 REFERENCES

[1] Makhoul, J. et al. (2000). Speech and Language Technologies for Audio Indexing and Retrieval: Advances and Applications. Proceedings of the IEEE, 88:1338-1353.

[2] Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. Journal of the Acoustic Society of America, vol. 58, No. 4, pp.880-883.

[3] Sakaguchi, S., Arai, T. and Murahara, Y. (2000). The effect of polarity of speech on human perception and data hiding as an application. ICASSP 2000.

[4] Greenberg, S. and Kingsbury, B. (1997). The modulation spectrogram: in pursuit of an invariant representation of speech. ICASSP 1997, pp. 1647- 1650.

[5] Kedem, B. (1986). Spectral analysis and discrimination by zero-crossing. Proceedings of IEEE, VOL. 74, No. 11, pp. 1477-1493.

[6] Kay, S. A zero-crossing based spectrum analyzer. (1986). IEEE Transaction on Acoustic, Speech and Signal Processing, vol. ASSP-34, No. 1, pp. 96-104.

[7] Sreenivas, T. and Niederjohn, R. (1992). Zero-crossing based spectral analysis and SVD spectral analysis for formant frequency estimation in noise. IEEE Transaction on Signal Processing, vol. 40, No. 2, pp. 282-293.

[8] Kim, D.S., Lee, S.Y. and Kil, R.M. (1999). Auditory processing of speech signal for robust speech recognition in real-world noise environment. IEEE Transaction on Speech and Audio Processing, 7, No. 1, pp. 59-69.

[9] De Mori, R., Moisa, L., Gemello, R., Mana, F. and Albesano, D. (2001). Augmenting standard speech recognition features with energy gravity centres. Computer Speech and Language, 15, pp. 341-354.

[10] Saunders, J. (1996). Real-Time Discrimination of Broadcast Speech/Music", Proc. ICASSP 1996, pp. 993-996.

[11] Scheirer, E. and Slaney, M. Construction and evaluation of a robust multifeature speech/music discriminator. ICASSP 97, vol. 2, pp. 1331-1334.

[12] Vintsiuk, T., Kulias, A. and Dys, A. (1982). A. economic presentation of acoustic signal in computer. ARSO-12, Institute of Cybernetics of AS USSR.

[13] Eickeler, S. and Rigoll, G. (2000). A novel error measure for the evaluation of video indexing systems. ICASSP 2000.

[14] Eickeler, S., Larson, M., Rüter, W. and Köhler, J., (2002) Creation of an Annotated German Broadcast Speech Database for Spoken Document Retrieval," to appear, Proceedings of LREC2002.