A Hybrid System for Audio Segmentation and Speech-endpoint Detection of Broadcast News

Maria Markaki[1], *Alexey Karpov*[2], *Elias Apostolopoulos*[1],
Maria Astrinaki[1], *Yannis Stylianou*[1], *Andrey Ronzhin*[2]
[1]Multimedia Informatics Lab, Computer Science Department, University of Crete (UoC), Greece
{mmarkaki,ilapost,astrin,yannis}@csd.uoc.gr
[2]St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences
{karpov,ronzhin}@iias.spb.su

Abstract

A hybrid speech/non-speech detector is proposed for the pre-processing of broadcast news. During the first stage speech/non-speech classification of uniform overlapping segments is performed. The accuracy in the detection of boundaries is determined by the degree of overlap of the audio segments and it is 250 ms in our case. Extracted speech segments are further processed on a frame basis using the entropy of the signal spectrum. Speech endpoint detection is accomplished with an accuracy of 10 ms. The combination of the two methods in one speech/non-speech detection system, exhibits the robustness and accuracy required for subsequent processing stages like broadcast speech transcription and speaker diarization.

1. Introduction

Automatic audio classification and segmentation is a research area of great interest in multimedia processing for automatic labeling and extraction of semantic information. In the case of broadcast audio recordings, pre-processing for speech/non-speech segmentation greatly improves subsequent tasks such as speaker change detection and clustering as well as speech transcription. Regarding speaker diarization systems, elimination of non-speech frames is more critical whereas for speech transcription accurate detection of speech is equally important.

In broadcast news, silence is usually reduced to a minimum and what mostly appears are other noises and music. Moreover, methods that work well on speech/music discrimination usually do not handle efficiently other non-speech classes commonly present in broadcast data such as environmental noises, moving cars, claps, crowd babble, etc.

Speech/non-speech segmentation can be formulated as a pattern recognition problem where the optimal features and the classifier built on them are application-dependent. Reviewing relevant past work, many approaches in the literature have examined various features and classifiers. MFCCs and SVMs have been extensively evaluated and seem to be among the most promising ones [1,2]. Furthermore, it has been shown that for successful audio segmentation and classification, the classification unit has to be a segment i.e. a sequence of frames rather than a single frame [1,2].

In this work we present a hybrid approach which combines a segment based classifier with a frame-based speech endpoint detector [3]. We use uniformly spaced overlapping audio segments of 500 ms length during the first classification stage. Mean and standard deviation of MFCCs have been used to parameterize every segment. We have also evaluated two different methods of spectrogram computation before MFCCs extraction. Classification is performed using SVMs [4]. During next stage, only segments characterized as speech are processed on a frame basis (10 ms). Spectrum entropy is the feature we use for the detection of silent frames within speech segments.

The organization of the paper is as follows: we review the segment-based speech/non-speech classification algorithm and the speech-endpoint detection method in section 2. In section 3 we describe experimental setup, the database and the experimental results. Finally in section 4 we present our conclusions.

2. Description of the method

2.1. Segment parameterization and classification

Mel-frequency cepstral coefficients are the most commonly used features in speech and speaker recognition systems. They also have been successfully applied in audio indexing tasks [1,2]. Here we extract 13th order MFCCs from audio frames of 25 ms with a frame rate of 10 ms, i.e. every 10 ms the signal is multiplied using a Hamming window of 25 ms duration. We perform critical-band analysis of the power spectrum with a set of triangular band-pass filters as usual. For comparison purposes, we also derive an auditory-like spectrogram by applying equal-loudness pre-emphasis and cube-root intensity-loudness compression according to Hermansky [5]. In each case, Mel-scale cepstral coefficients are computed every 10 ms from the filterbank outputs. We define each segment as a sequence of 50 frames of 10ms each. We estimate the mean and standard deviation of MFCCs over 50 frames, resulting in a 26-element feature vector per segment. We extract evenly spaced overlapping segments every 25 frames (250 ms overlap) for the test dataset whereas for the training dataset segments are extracted every 5 frames (for maximizing training data).

Support vector machines (SVMs) are used for the classification of segments. We have used SVM^{light} [4] with a Radial-Basis-Functions kernel – all the other parameters have been set to default values. We also define an hierarchy of classes similar to [2] for resolving conflicts that arise due to the overlap of segments: frames are classified as non-speech if they are part of any segment that was classified as non-speech; otherwise, they are classified as speech.

2.2. Spectral entropy-based speech detector

The speech detection method is based on calculation of the information entropy of the signal spectrum as the measure of uncertainty or disorder in a given distribution [6]. The distinction between entropy for speech segments and entropy for background noise is used for speech endpoint detection. Such criterion is less sensitive to the variations of the signal amplitude than the energy-based methods. The method is a modification of the speech detection approach proposed by J.-L. Shen [7] and includes new levels into the analysis of speech signal (Figure 1).

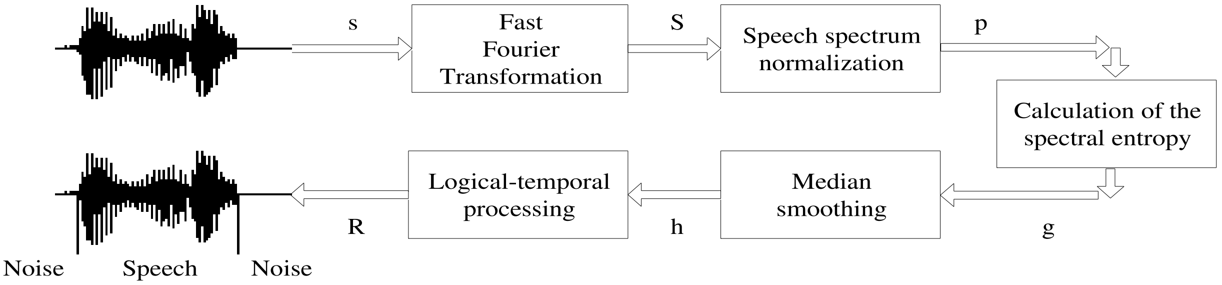
s
Fast Fourier Transformation
S
Speech spectrum normalization
p
Calculation of the spectral entropy
g
Median smoothing
h
Logical-temporal processing
R
Noise
Speech
Noise

Fig. 1. The algorithm for speech detection based on analysis of the entropy of signal spectrum

The audio signal is divided into short segments with duration 11.6 ms each with overlapping 25%. Short-time signal spectrum is computed using FFT, and normalization of the calculated spectrum over all frequency components is fulfilled giving the probability density function p_i. Acceptable values of probability density function are upper and lower bounded. This restriction allows us to exclude noises concentrated in a narrow band as well as noises approximately equally distributed among the frequency components (for instance, white noise). Thus:

$$p_i = 0, \text{ if } p_i < \delta_2 \text{ or } p_i > \delta_1 \qquad (1)$$

where δ_1 and δ_2 are the upper and lower values of probability density, respectively. They have been experimentally determined to be $\delta_1 = 0.3$ and $\delta_2 = 0.01$. At the next stage the information spectral entropy *h* is estimated, and median smoothing in a window of 5-9 segments is applied. Finally, a logical-temporal processing of *h* (Figure 2) takes into account the possible durations of speech and non-speech fragments.

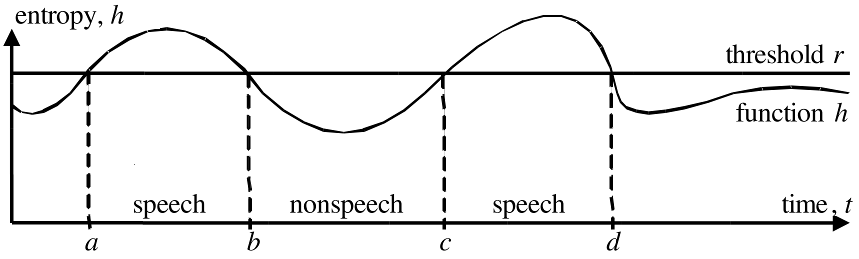

Fig. 2. Logical-temporal processing of the spectral entropy function

An adaptive threshold r for the detection of speech endpoints is calculated as follows:

$$r = \left[\frac{\max(h) - \min(h)}{2} + \min(h)\right] * m \qquad (2)$$

where μ is a coefficient empirically chosen depending on the recording conditions. Employing the adaptive threshold we can obtain alternate speech and non-speech regions given the function *h* and apply two criteria to process: (1) R – minimal duration of a speech fragment in a phrase; (2) S - maximal duration of a non-speech fragment in a phrase. These criteria values were experimentally determined taking into account that a human cannot produce very short speech fragments as well as that there are always some pauses in speech (for instance, before explosive consonants). So if the number of consecutive speech segments is greater than *R* and non-speech interval between them is shorter than *S* then all these segments are considered belonging to speech class. Such logical-temporal processing is applied iteratively to the whole spectral entropy function automatically segmented for speech/non-speech portions.

3. Experiments and Results

We tested the algorithms described in section (2) on audio data collected from Greek TV programs (TV++) and music CDs. Speech data consists of broadcast news and TV shows recorded in different conditions such as studios or outdoors; also, some of the speech data have been transmitted over telephone channels.

Non-speech data consists of music (25%), outdoors noise (moving cars, crowd noise, etc), claps, and very noisy unintelligible speech due to many speakers talking simultaneously (speech babble). Music content consists of the audio signals at the beginning and the end of TV shows as well as songs from music CDs. Audio data are all mono channel and 16 bit per sample, with 16 kHz sampling frequency. The database has been manually segmented and labeled at Computer Science Department, UoC. Speech signals have been partitioned into 30 minutes for training and 90 minutes for testing.

3.1. Speech / non-speech classification results

We evaluate system performance using the detection error trade-off curve (DET) [8]. DET plot clearly presents detection performance tradeoff between false rejection rate (or speech miss probability) and false acceptance rate (or false alarm probability). Detection error probabilities are plotted on a nonlinear scale which transforms them by mapping to their corresponding Gaussian deviate. Thus DET curves are straight lines when the underlying distributions are Gaussian [8].

We also report the minimum value of the detection cost function for each detection error trade-off curve according to [8]. For the speech/non-speech segment-based classification, the target is speech class having prior probability $P_{t\arg et} = 50\%$ in our data set.

Here the costs of miss and false alarm probabilities are considered equally important ($C_{miss} = C_{false} = 1$) although they actually depend on the task. For speaker and language recognition $C_{false} > C_{miss}$, i.e. we should accurately reject non-speech audio (low false alarm probability) whereas speech miss probability is less important. For speech transcription on the other hand $C_{false} < C_{miss}$, i.e. accurate detection of speech is rather more important.

The minimum value of the detection cost function (DCF) for the DET curve [8] then, is:

$$DCF_{opt} = \min(C_{miss} * P_{miss} * P_{t\arg et} + C_{false} * P_{false} * (1 - P_{t\arg et})) \tag{3}$$

In the case of common MFCC features, $DCF_{opt} = 9.54\%$ and corresponds to $P_{miss_{opt}} = 6.24\%$ and $P_{false_{opt}} = 12.84\%$. For the case of MFCC features extracted after loudness equalization and cube root compression, a remarkable improvement in all aspects is noticed: $DCF_{opt} = 4.96\%$, $P_{miss_{opt}} = 4.07\%$ and $P_{false_{opt}} = 5.84\%$.

Another commonly used measure of accuracy is the EER (Equal Error Rate) which corresponds to the decision threshold θ_{EER} at which false rejection rate (P_{miss}) equals false acceptance rate (P_{false}). Since P_{miss} and P_{false} are discrete, we set:

$$\theta_{EER} = \arg\min_q \left| P_{miss}(q) - P_{false}(q) \right| \tag{4}$$

and

$$EER(q) = \frac{P_{miss}(q) + P_{false}(q)}{2} \tag{5}$$

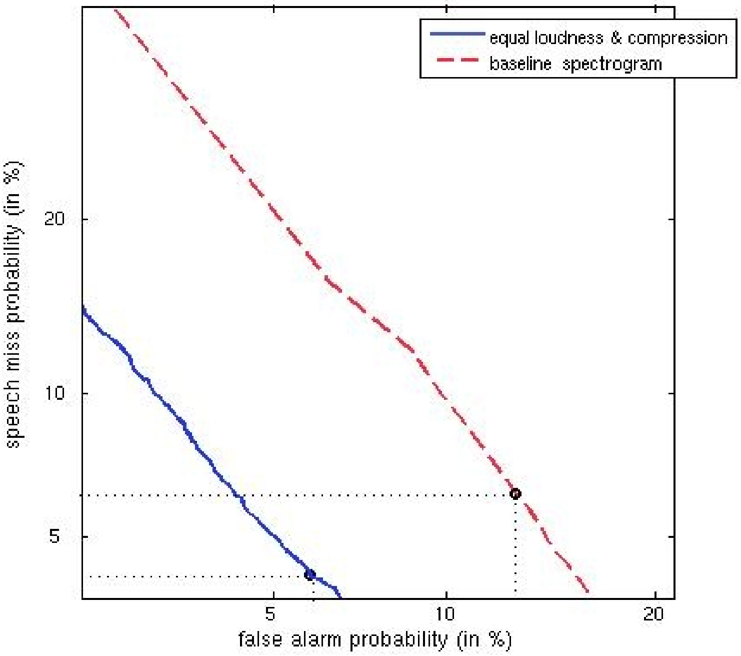

Figure 3: DET curves for speech/non-speech segment-based classification. Mean and variance of MFCCs are computed over each segment, with (solid line) or without (dashed line) equal-loudness pre-emphasis and cube-root intensity-loudness compression [5]. The minimal values of the corresponding detection cost functions (DCF) are also presented (circles).

We report in Table 1 the results for the speech/non-speech segment-based classification and present in Figure 3 the corresponding DET curves. Since in this case $P_{t\,\mathrm{arg}et} = 50\%$ and $C_{miss} = C_{false} = 1$, both values of EER and DCF$_{\mathrm{opt}}$ are quite close. MFCC features extracted after loudness equalization and compression are clearly superior according to EER, too.

Table 1: Speech/non-speech segment based classification results

System	DCF$_{\mathrm{opt}}$	P$_{\mathrm{miss}}$	P$_{\mathrm{false}}$	EER
MFCCs baseline	9.54%	6.24%	12.84%	9.91%
equal loudness+compression	4.96%	4.07%	5.84%	5.01%

3.2. Speech endpoint detection results

Audio segments classified as speech at the first detection stage are further processed using the entropy based method for speech-endpoint detection with 10 ms accuracy (after rounding). This is a pre-processing step required for subsequent broadcast speech transcription. In this case, the total number of silence frames is much lower than the total number of speech frames: prior probability of speech class is $P_{t\,\mathrm{arg}et} = 88.96\%$ for our dataset where speech is the target. If the costs of miss and false alarm probabilities are considered of equal importance, then the minimum value of the detection cost function (DCF) for the DET curve is $\Delta X\Phi_{o\pi\tau} = 6.47\%$ corresponding to $\Pi_{\mu\iota\sigma\sigma_{o\pi\tau}} = 4.48\%$ and $\Pi_{\phi\alpha\lambda\sigma\epsilon_{o\pi\tau}} = 22.52\%$. We report in Table 2 the results for speech/silence classification and present in Figure 4 the corresponding DET curve. We can see that EER is significantly higher than DCF_{opt} in this case since it doesn't take into account the highly unequal prior probabilities of speech and silence.

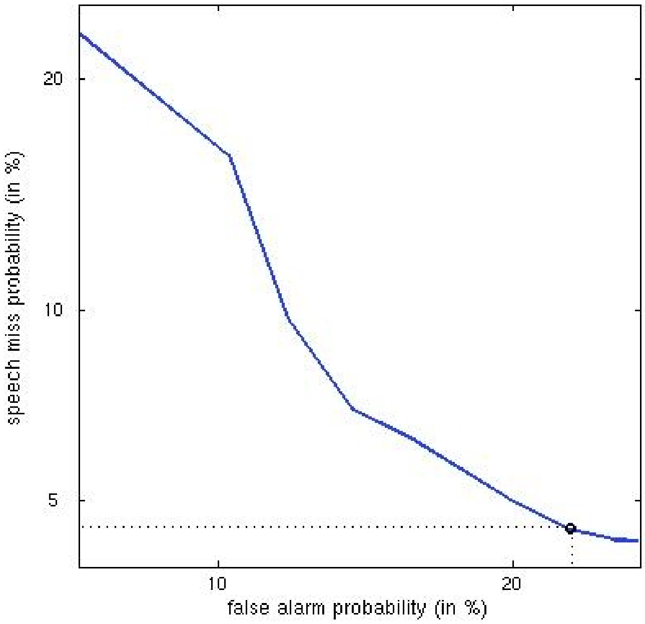

Figure 4: DET curve for speech endpoint detection with 10 ms accuracy applied onto extracted speech segments. The minimal value of the corresponding detection cost function (DCF) is presented as circle.

Table 2: Speech/silence classification results based on spectrum entropy

DCF_{opt}	P_{miss}	P_{false}	EER
6.47%	4.48%	22.52%	10.83%

4. Conclusions

In this paper we have applied a two-stage speech detection system. During the first stage, segment-based speech/non-speech classification is performed based on MFCC features and Support Vector Machines within 250 ms accuracy. An improvement is reported if we use loudness equalization and cube-root compression to the power spectrogram after critical-band analysis. Extracted speech segments are further processed through an entropy-based method for speech-endpoint detection within 10 ms accuracy. The proposed system can successfully address the two-fold requirement for robustness and accuracy during the pre-processing stages preceding broadcast speech transcription or speaker diarization.

Acknowledgements

This work has been supported by the General Secretariat of Research and Technology (GGET) in Greece and Russian Foundation for Basic Research in the framework of the project # 07-07-00073-a. The collaborative research was part of the PhD exchange program of the SIMILAR Network of Excellence project # FP6-507609.

References

1. *L. Lu, H.J. Zhang, Stan Li.* Content-based audio classification and segmentation by using support vector machines. Multimedia Systems 8: 482-492, 2003.

2. *H. Aronowitz.* Segmental modeling for audio segmentation. Proc. ICASSP 2007, Hawaii, USA, 2007.

3. *A. Karpov.* A robust method for determination of boundaries of speech on the basis of spectral entropy. Artificial Intelligence Journal. Donetsk, Vol.4. pp. 607-613, 2004.

4. *T. Joachims*. Making large-scale SVM learning practical. In Advances in Kernel Methods – Support Vector Learning, MIT-Press, 1999.

5. H. Hermansky, B. Hanson, H. Wakita. Perceptually based linear predictive analysis of speech. Proc. ICASSP 1985, pp. 509-512, 1985.

6. *J. Ajmera, I. McCowan, H. Bourlard.* Speech/music segmentation using entropy and dynamism features in a HMM classification framework. Speech Communication, № 40, pp. 351-363, 2003.

7. *J.-L. Shen, J.-W. Hung, L.-S. Lee.* Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments. Proc. ICSLP 1998, Sydney, Australia, paper 0232, 1998.

8. The NIST Year 2004 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2004/