Heuristic and Statistical Methods for Speech/Non-speech Detector Design

Michal Prcín and Luděk Müller

University of West Bohemia, Department of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic
mprcin@kky.zcu.cz
muller@kky.zcu.cz

Abstract. Speech/non-speech (S/NS) detection plays the important role for automatic speech recognition (ASR) system, especially in the case of isolated words or commands recognition. Even in continuous speech a S/NS decision can be made at the beginning and at the end of a sequence resulting in a "sleep mode" of the speech recognizer during the silence and in a reduction of computation demands. It is very difficult, however, to precisely locate the endpoints of the input utterance because of unpredictable background noise. In the proposed method in this paper, we make use of the advantages of two approaches (i.e. to try to find the best set of heuristic features and apply a statistical induction method) for the best S/NS decision.

1 Introduction

S/NS detection is an ubiquitous problem in speech processing. It consists of the classification of two clearly distinct signal conditions during speech recording: periods where the speech signal is present and pauses with background noise only. Generally, it is a very difficult task because speech waveform can be obtained in various environments (street, car, factory, office, etc.). The non-speech regions can involve sounds such as the opening or closing of a door, coughing, background music, vibrations from an engine, and so on. While significant work has been done on this problem, most efforts have attempted to extract speech frames from data containing only one specific signal type. Unfortunately, when such an algorithm is applied to separate speech from data containing different specific non-speech signals, results are generally poor.

The remainder of this paper is organized as follows. Section 2 depicts methods used in speech detection. Section 3 describes our approaches to find a set of optimal features. Section 4 includes comparative tests with a different feature set and Section 5 concludes the paper.

2 Comparison of Heuristic and Statistical Methods

Almost all algorithms of speech detection can be split into two classes. Methods of the first class are heuristic methods while methods of the second one are based on a statistical approach. The advantage of heuristic methods is their effort to find a set of optimal features for the speech/non-speech classification. Their drawback is that they use thresholds that are set heuristically (i.e. their value estimate is based on the experience of a human expert). These predefined thresholds are then used to classify a signal frame as "speech" or "non-speech". The statistical methods do not suffer from this shortcoming because their threshold setting is done more precisely by using training data and applying a statistical induction method. The disadvantage of statistical methods referred to in literature consists in the use of a feature set based on some kind of a standard speech parameterization (e.g. MFCC or PLP). We decided to use the advantages of both approaches (i.e. to try to find the best set of heuristic features and apply a statistical induction method for the speech/non-speech classifier optimal setting). The aim is to develop a classification algorithm that will best classify each audio signal frame as a speech or non-speech frame among all classifiers working with any heuristic feature set.

3 Selection of Optimal Features

First of all, features commonly used in a heuristic approach were collected creating a large basic feature set (feature vector) and then statistical speech and non-speech models were trained using EM algorithm. The training data were represented by a speech/non-speech signal transformed (i.e. parameterized) into feature vector sequences. The resulting basic feature vector comprised the following coefficients: short-time zero crossing rate (STZCR), short-time energy (STE), short-time intensity (STI), entropy (H), EE-Feature (Energy-Entropy feature), STZCR to STE ratio (ZER), product of STZCR and energy (ZEP), RMS to mean ratio (RMR), normalized correlation coefficient for lag 1 (NCC1), and periodicity of the autocorrelation functions. We also trained the classifier using a set of features commonly used in speech recognition (MFCC, PLP (+ RASTA)). Furthermore, the basic feature set was modified in several ways (for example by considering log energy instead energy, various modifications of autocorrelation functions periodicity, signal normalization, etc.).

3.1 Parameterization with Features Used in Heuristic Approach

The feature vector of this parameterization contained 11 features that describe a given frame. These features were selected from features used in [1], [2], and [4].

Computation of individual features:

1. Short-time energy
   For each frame i, the energy E_i of the frame is obtained by the sum of squares:
$$E_i = \sum_{n=1}^{N} s_n^2 \ , \tag{1}$$

where s_n and N are nth signal and the total number of signals in frame i respectively.
Instead the E_i also the LE_i can be used:

$$LE_i = \log E_i \ . \tag{2}$$

2. Entropy
   For each frame i, signal within it is converted from time domain to frequency domain via the FFT equation:

$$X(\omega) = \sum_{n=-\infty}^{\infty} s_n e^{-j\omega n} \ . \tag{3}$$

   Then the probability density function (pdf) for spectrum can be estimated by normalizing the frequency components:

$$p_i = \frac{s(f_i)}{\sum_{k=1}^{M} s(f_k)} \ , \tag{4}$$

   where $s(f_i)$ is the spectral energy of the frequency component f_i, p_i is the corresponding probability density, and M is the total number of frequency components in FFT. Some heuristic constraints are used to improve the discriminability of the pdf between speech and non-speech signals and only the frequency from 250Hz to 3750Hz are considered. The limitation of the range of frequency is just because the range covers most of the frequency components of human speech signals. Therefore, we define the equations (5) and (6) below:

$$s(f_i) = 0, \quad \text{if } f_i \leq 250Hz \text{ or } f_i \geq 3750Hz \ , \tag{5}$$

$$p_i = 0, \quad \text{if } p_i \geq 0.9 \ . \tag{6}$$

   The negative entropy H_i of frame i can be defined as:

$$H_i = -\sum_{j=1}^{M} p_j \log_2 p_j \ . \tag{7}$$

3. Energy-Entropy Feature
   EE-Feature is obtained as:

$$M_i = (E_i - C_E) \cdot (H_i - C_H) \ , \tag{8}$$

$$EE - Feature_i = \sqrt{1 + |M_i|} \ , \tag{9}$$

   where C_E a C_H denote the average energy and the entropy of the first 10 frames, respectively. Because both energy and entropy have their limitations, the blind spots in either energy or entropy, or both, can be cancelled by their multiplication. In other words, energy covers the case that was failed in entropy: babble and background music in speaker utterance; whereas the entropy covers the case that was failed in energy: non-stationary noise which belongs to mechanical sounds.

4. Short-time Intensity
For each frame i, the intensity M_i of the frame is obtained by the sum of absolute values of s_n:

$$M_i = \sum_{n=1}^{N} |s_n| \ , \tag{10}$$

where s_n and N are nth signal and the total number of signals in frame i respectively.

5. Short-time zero crossing rate (STZCR)
The equation for computing the number of zero crossings in a block can be obtained as:

$$Z_i = \sum_{n=0}^{N-1} |\mathrm{sgn} s_n - \mathrm{sgn} s_{n-1}| w_{N-1-n} \ , \tag{11}$$

where

$$\mathrm{sgn} s_n = \left\langle \begin{array}{l} 1 \ \text{ pro } s_n \geq 0 \\ -1 \text{ pro } s_n < 0 \end{array} \right. \tag{12}$$

and $w_N - 1 - n$ is a rectangle window.
STZCR can be computed as the number of zero crossings per second, using the formula

$$STZCR_i = \frac{Z_i . f_s}{N} \ , \tag{13}$$

where Z_i is the number of zero crossings in a block with N samples and f_s is the sampling frequency.

6. ZER

$$ZER_i = \frac{STZCR_i}{E_i} \ . \tag{14}$$

7. ZEP

$$ZEP_i = STZCR_i \cdot E_i \ . \tag{15}$$

8. RMR
The ratio of the RMS to mean value is evaluated as RMR:

$$RMR_i = \sum_{n=0}^{N-1} \frac{\sqrt{\frac{1}{N} s_n^2}}{\left|\frac{1}{N} s_n\right|} \ . \tag{16}$$

9. Normalized Correlation Coefficient for lag 1 (NCC1)
NCC1 is obtained by using the formula:

$$NCC1_i = \frac{\sum\limits_{n=2}^{N} s_n \cdot s_{n-1}}{\sqrt{\sum\limits_{n=2}^{N} s_n^2 \sum\limits_{n=2}^{N} s_{n-1}^2}} \ . \tag{17}$$

10. M1
The autocorrelation coefficient $R(J)$ is calculated as

$$R\left(J\right)_i = \sum_{n=J+1}^{N} s_n \cdot s_{n-J} \ , \tag{18}$$

$M1$ is determined by noting the value of J for which $R(J)$ is a maximum for values of J greater than 15.
11. M2
Similarly, $M2$ is founded by considering $J = 15 + M1$. The same equation is used:

$$R\left(J\right)_i = \sum_{n=J+1+M1}^{N} s_n \cdot s_{n-J} \ . \tag{19}$$

Furthermore, the basic feature set was modified in several ways (for example by considering log energy instead energy, various modifications of autocorrelation functions periodicity, signal normalization, etc.).

In the following, presented MPP parameterizations are defined:

- MPP (ble) - the feature vector consisted of 11 coefficients mentioned in Section 3.1. For energy computing the equation (1) was used.
- MPP_D_A (sle) - to the MPP (ble) coefficients delta and accelerations coefficients were added.
- MPP (sle) - the feature vector had 11 coefficients but the equation (2) instead (1) was applied and in others equations log energy instead energy was used.
- MPP (8p) - the feature vector was comprised 8 coefficients described in Section 3.1. The equation (2) instead (1) was employed and feature vector did not contain neither energy nor intensity. Features $M1$ and $M2$ were replaced by $|2 \cdot M1 - M2|$.
- MPP (jebl) - the feature vector had 1 coefficient computed according to the equation (1).

3.2 Features Commonly Used in Speech Recognition

We also trained the classifier using a set of features commonly used in speech recognition (MFCC, PLP (+ RASTA))[3]. Firstly, we obtained a feature vector with 27 MFCC_D_A (Mel-Frequency Cepstral Coefficients Delta Acceleration), that consisted of 9 static MFCC, 9 delta MFCC and 9 acceleration MFCC.

The next parameterization was PLP_D_A (Perceptual Linear Predictive Analysis Delta Acceleration). The feature vector also described 27 coefficients for each given frame: 8 static PLP + energy, 9 delta and 9 acceleration coefficients.

Finally, the third parameterization RASTA_D_A (RelAtive SpecTrAl Delta Acceleration) consisted of 9 RASTA coefficients, 9 delta PLP and 9 acceleration

PLP. 9 RASTA coefficients were acquired from 8 static PLP coefficients and 1 energy coefficient by applying the RASTA filter with a transfer function:

$$H(z) = 0.1z^4 \cdot \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \ . \tag{20}$$

Furthermore, we decided to join all presented features in the subsection 3.2 into a total vector of parameterization TOGETHER (MPP+MFCC+PLP+RASTA_D_A). This total vector had 53 coefficients for each frame (8 MPP, 9 static MFCC, 9 static PLP and 27 RASTA_D_A coefficients).

4 S/NS Detection Experiments

In order to find the best parameterization technique (the best feature set) we had to perform many experiments for various modifications of the basic feature set. The experiments were accomplished on a set of waveforms (files) that were cut out from speech utterances ($\mu - law$ format, $8kHz$ sample rate) obtained from phone calls about information on train connections. The waveforms were divided into two classes. The first class contained non-speech (low-level noise) files. 379 waveforms were used for training and 22 files for testing. The second class consisted of waveforms containing only speech (1324 for training and 100 for testing).

Two statistical models were trained for each kind of parameterization. The first model was a model of speech (S_CH), which was trained with waveforms containing only speech. The second one (i.e. model of non-speech (N_S)) was trained on non-speech (low-level noise) waveforms. This pair of model represents one speech/non-speech detector and was trained for various number of components of Gaussian mixture. Each feature vector is evaluated by S_CH model and N_S model. If the probability assigned by the S_CH model is greater than the probability assigned by the N_S model, given frame is detected as speech and vice versa.

The accuracy of each proposed speech/non-speech detector was then investigated. The tests accomplished so far were done on waveforms containing only clean speech or only low-level noise (silence). As can be seen from Figure 1, the decision whether given frame is speech or non-speech, is performed by using S/NS detection network.

Many tests were carried out for various parameterizations (feature set). Table 1 lists the S/NS detection accuracy on 5 speech utterances with quiet, (low-level) volume. Each parameterization (e.g. PLP_D_A) was trained for various number of components of Gaussian mixture. Such number of components was selected from each parameterization (e.g. from all PLP_D_A parameterizations) that best classified all tested waveforms. The PLP_D_A parameterization had the best results for the number components of Gaussian mixture 26. As mentioned earlier, the tests were accomplished only on waveforms with clean speech or low-level noise (without speech) and each pair of model representing one speech/non-speech detector were used for the S/NS decision.

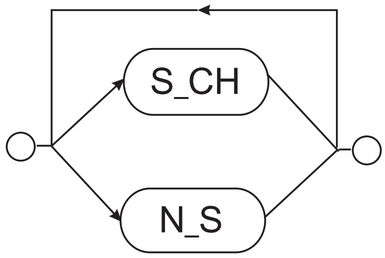

Fig. 1. Recognition network of the S/NS detection

Table 1. A comparison of the best selected parameterizations on 5 speech waveforms

SPEECH WAVEFORMS:	X0054	X0436	X0552	X0681	X0686	SPEECH	NOISE
Number of frames::	109	34	104	84	34	365	0
Parameterizations							
MPP (ble) mix 22	78/31	0/34	3/101	23/61	0/34	104	261
MPP_D_A (sle) mix 18	21/88	0/34	0/104	22/62	0/34	43	322
MPP (sle) mix 8	109/0	0/34	104/0	79/5	6/28	298	67
MPP (8p) mix 20	109/0	0/34	104/0	29/55	8/26	213	152
MPP (jebl) mix 32	72/37	0/34	23/81	30/54	11/23	173	192
MFCC_D_A mix 10	109/0	2/32	101/3	48/36	9/25	269	96
PLP_D_A mix 26	109/0	4/30	91/13	50/34	34/0	288	77
RASTA_D_A mix 22	109/0	29/5	104/0	47/37	34/0	323	42
TOGETHER mix 10	109/0	5/29	104/0	46/38	32/2	296	69

For example, row 10 in Table 1 describes PLP_D_A mix 26. It means that the feature vector with PLP coefficients (8 static PLP + energy + 9 delta PLP + 9 acceleration PLP) was used for the S/NS detection. The number components of Gaussian mixture is 26. The testing speech waveform X0552 consists of 104 frames. 91 frames were detected as speech and 13 frames as non-speech. The column "SPEECH" describes the sum of frames that were detected as speech (a maximum value is 365). It can be seen from Table 1 that RASTA_D_A mix 12, PLP_D_A mix 26 and MPP (sle) mix 8 are the best parameterizations for S/NS detection. Since the results are very similar and PLP coefficients for speech recognition are used in our ASR system, we decided to apply the feature vector with PLP coefficients to the S/NS decision too. The advantage of this choice is, that the same feature vector is first introduced to the S/NS detector and then for speech recognition.

5 Conclusions

This paper has shown a statistical approach to S/NS detection together with an investigation of several (heuristic) feature sets. The parameters of the models (speech and non-speech) were set by applying the statistical induction method.

The various parameterizations were tested for finding the best features set for the speech/non-speech decision in the utterance. When the speech waveforms were louder or waveforms were normalized, each of the proposed and presented parameterizations successfully detected waveforms as speech. However, all parameterizations had problems on waveforms with low-level speech (see in Table 1). All parameterizations infallibly detected the noise in low-level noise utterances.

6 Acknowledgment

This work was supported by the Ministry of Education of the Czech Republic, project no. FRVS 966/2002

References

1. Yang, C-H., Hsieh, M-S.: Robust Endpoint Detection for In-Car Speech Recognition,ICSLP 2000, Beijing, Paper Number 251
2. Zhu, J., Chen, F.: The Analysis and application of a New Endpoint Detection Method Base on Distance of Autocorrelated Similarity, ESCA, Eurospeech99, Budapest, Hungary. ISSN 1018-4074, Page 105-108
3. Karray, L., Monné, J.: Robust Speech/non-speech Detection in Adverse Conditions Based on Noise and Speech Statistics,ICSLP'98, Sydney, Paper Number 430
4. Tessama, T.: New Methods for the Detection of Voiced, Unvoiced, Silent, Transition Regions in Speech Signals, Advances in Modelling & Analysis, B, AMSE Press(France), vol.31, No.3, 1994, pp 1-10 Received Oct. 19, 1993