THE METHOD FOR RECOGNITION THE ACTUAL EMOTIONAL STATE OF SPEAKER

Miroslav Staněk

Doctoral Degree Programme (1), FEEC BUT

E-mail: xstane03@stud.feec.vutbr.cz

Supervised by: Milan Sigmund

E-mail: sigmund@feec.vutbr.cz

Abstract: This paper deals with the differences of important points in LPC spectrum depending on emotional state for Czech vowels. LPC spectrum contains antiformats- the local minima, formants- the local peaks and its points of inflection. The position of all important points depends on actual emotional state of speaker. Significant differences in distribution were obtained for each important point of LPC spectrum within the emotional states of speaker database for F2-F3 and the best for F1-F2 band. Most similar shapes of histograms were achieved for alcohol intoxication and happiness. Ordinary values of relevant histogram parameters for each important point within the emotional states of speaker database are a matter of next statistical investigation.

Keywords: speech processing, emotion recognition, formant analysis

1. INTRODUCTION

Human is being with the most diverse emotions on the whole World. The actual emotional state can be presented by oral way (speech) or by body language. This paper is oriented on emotions occupied in speech signal and its processing for determining them. It is also necessary to mentioned that many ways to determining the emotion state of speaker exist.

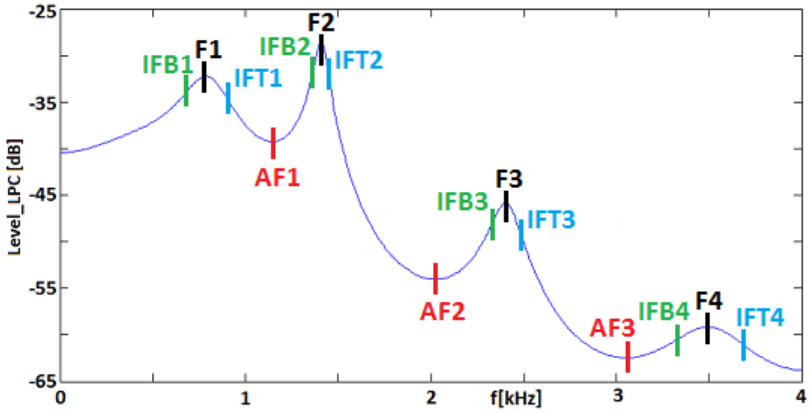

Figure 1: LPC spectrum of vowel /a/ and its important points.

The investigation of this topic is important in e.g. interaction human-robot, psychical diseases or alcohol detection, real voice synthesis etc. In references [1], [2], [3], [4], [5] are listed mostly used speech features for emotional detection such as pitch, short time signal energy, duration, line spectral frequency etc. The inverse process of speech signal speech processing is speech synthesis. For best results of real voice synthesized by machine are used different models of adaption techniques, e.g. Gaussian Mixture Models and so on. More information about modelling the emotional states for speech synthesis and emotion recognition are listed in references [6].

As it was written, this paper is focused on parameters (level and frequency) of important points in LPC spectrum for Czech vowels. The example of /a/ vowel LPC spectrum with marked important points is in Figure 1 where formants are F, antiformants are AF, the top point of inflection of particular formant is IFT and bottom points of inflection are labelled as IFB.

2. EXTRACTING SPEECH FEATURES

Next section of this paper describes the whole processing of recorded speech signal. Recorded fluent speech is sampled by sampling frequency 8 kHz and it is divided to short segments. Each segment lasts 20 ms. Because of the time duration and sampling frequency the size of each segment is 160 samples. If some segment contains less than 160 samples is fulfilled by zeros up to this length. After the record dividing each segment is processed separately in order.

Current segment is normalized- the signal amplitude reaches on each segment values from -1 to +1, and LPC spectrum of current segment is calculated. Generated LPC spectrum is derived for the first time. By the first derivation the local extremes (formants and antiformants) are searched. In general can be said that the first two formants determine the spoken phoneme and the last two identify the speaker. If formant F1 and F2 lie in common frequency band the relevant vowel is detected on the current segment and the segment is marked as useful. In the other case the segment is marked as useless and next segment is going to be processed. Common formant frequencies for Czech vowels are written in Table1 [7].

Vowel	F1 [Hz]	F2 [Hz]	F3 [Hz]
/a/, /á/	700 – 1100	1100 - 1500	2500 – 3000
/e/, /é/	480 – 700	1560 - 2100	2500 – 3000
/i/, /í/	300 – 500	2000 - 2800	2600 – 3500
/o/, /o/	500 – 700	850 - 1200	2500 – 3000
/u/, /ú/	300 – 500	600 - 1000	2400 - 2900

Table 1: Common formant frequency bands for Czech vowels.

If the current segment is marked as useful the parameters of other important points in LPC spectrum are calculated. By the first derivation are also determined level and frequency of antiformants. The second derivation is used for searching the points of inflection and their parameters. Figure 2 shows the derivations of LPC spectrum illustrated in Figure 1. In Figure 2 are marked all important points in LPC spectrum by 4 kinds of markers. The second points of inflection are not marked for better illustration. These points lie near the second formant F2 in intersections of second derivation and zero level (see Figure 2).

Extracted parameters of useful segments for each vowel are saved for further processing. After the vowel recognition over the whole record extracted data are checked by retroactive algorithm based on the possible phoneme duration which is at least 40 ms. By this algorithm the differences in order number between neighbouring useful segments are checked. If the difference is bigger than 2 for previous and next extracted useful segment, the current segment is marked as useless. If the difference in order number is 1 at least for one neighbouring useful segment the current segment is still marked as useful and the missed useful segment is reanalyzed. By this algorithm the false detected segments has been reduced at average of 38.3%. The reduction of false detected segment is necessary because of the relevance of further results.

Extracted useful data of each vowel are further processed in statistical way. The used method is very fast and it has not been published yet. It is based on the changes of important point relative

position. Of course, the changes are caused in level and frequency criterion in addiction on actual emotional state of speaker. But this paper is oriented only on changes caused in frequency domain.

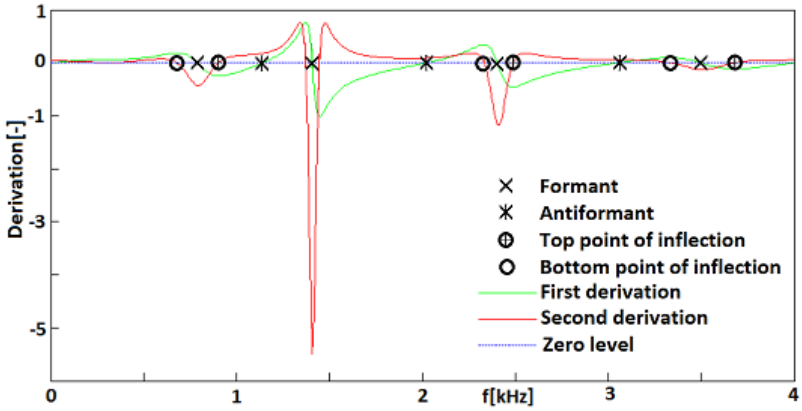

Figure 2: The derivations of LPC spectrum of vowel /a/ and its important points.

For investigation of changes in frequency domain are observed three frequency bands separately bounded by consecutive formants. Between both bounding formats lie lower formant top point of inflection, antiformant and bottom point of inflection for higher formant. It has to be noticed that the bottom point of inflection IFB1 and the top point of inflection IFT4 are not observed because they lie out of investigated bands.

Relative position *length* of important point in actual frequency interval is calculated as

$$length_X[\%] = \frac{Xf - A}{B - A} \cdot 100\,, \qquad (1)$$

where X is the current important point, Xf is its frequency, A is frequency of lower formant and B is higher formant frequency in observed interval. All calculated relative positions are further displayed by relevant histograms for better observation of the differences between emotional states.

3. EXPERIMENTAL RESULTS

For one speaker, the voluntary theatre actor, has been recorded emotional states like normal mood, anger, sadness, happiness and alcohol intoxication. All of these recorded emotional states are simulated.

From relevant records the useful speech features has been extracted for all vowels. In this section are illustrated only results obtained for /a/ vowel.

Results obtained by speech signal analysing are statistically processed for better changes illustration. Each investigated frequency band contains three histograms of extracted results. For better resolution are colours of generated histograms within obtained frequency band different. The lower formant top point of inflection is green, histogram of antiformant is red and the low bottom point of inflection of higher formant is blue.

In Figure 3 are illustrated histograms obtained for simulated emotional states in formant bands F1-F2 and F2-F3. All extracted histograms in Figure 3 are for vowel /a/ in the fluent speech where the same text has been spoken for all emotional states.

From Figure 3 is obvious the differences in histogram shapes for individual emotional state exist. For normal state of speaker are typical narrow histograms for all important points and high rate of relative position. Anger has also high rate of relative positions of important points but histograms are wider than for normal mood. The centre of histogram for the first antiformant is shifted from 55% (normal mood) to 40%. The histogram of the second antiformant has absolutely different

shape which is similar to histogram shape of bottom point of inflection of higher formant. Modus of histogram for AF2 in anger has relative position 60%

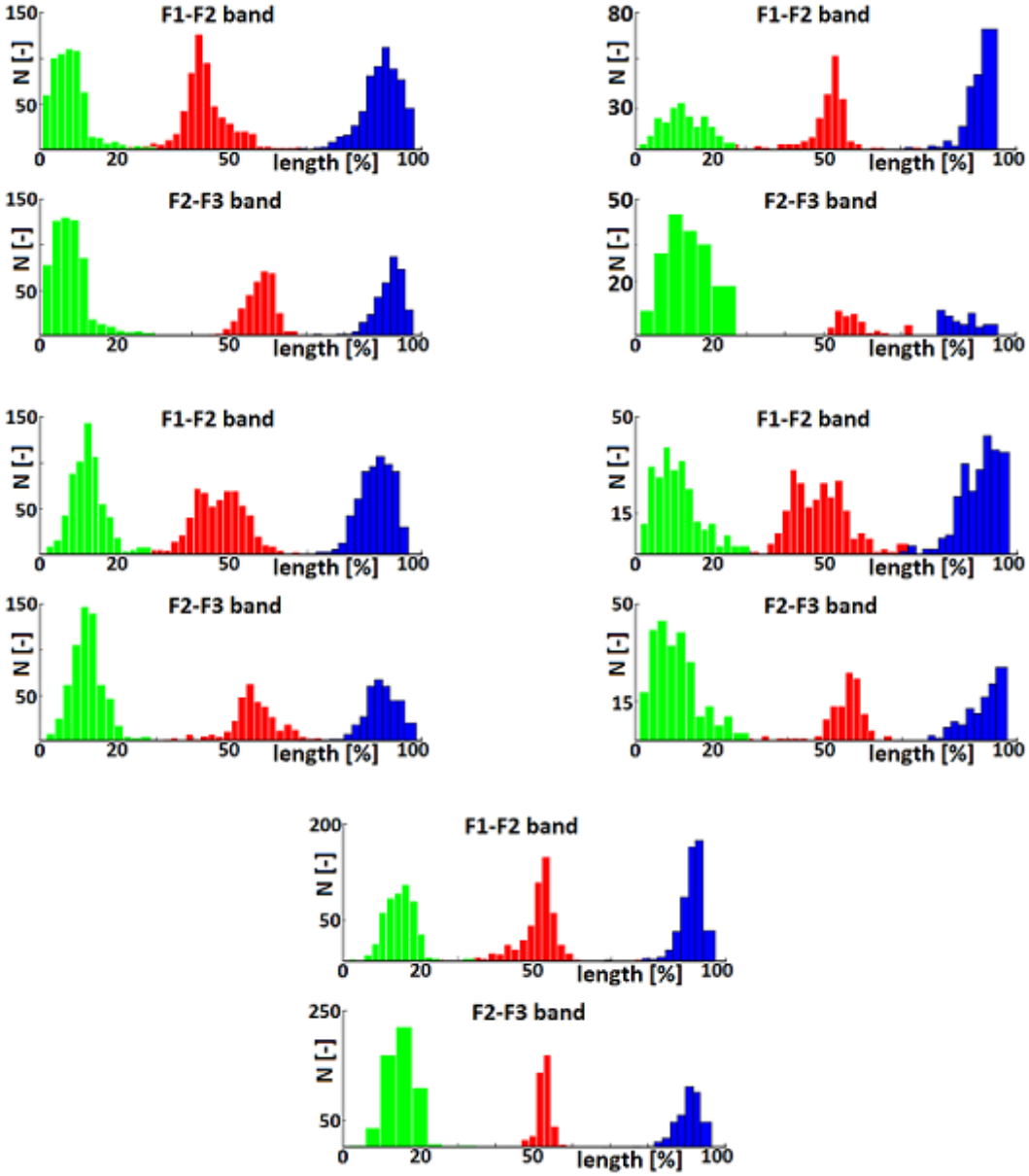

Figure 3: Relative positions of important points in F1-F2 and F2-F3 bands. From left to right, from up to down: anger, sadness, happiness, alcohol intoxication and normal mood.

For sadness state of speaker is typical less rate of each relative position. The least rate has AF2 and IFB3. This is caused because the third formant F3 is not generated by vocal tract in sadness. In the other case the states for happiness and alcohol intoxication are similar. Both states have wide histograms for each important point. The shapes of antiformants are more a less similar to normal (Gaussian) distribution. The difference between these moods is in the shape of histogram for points of inflection. For happiness the shape of histogram for both points of inflection has Ricean distribution and for alcohol intoxication is the shape of same histograms similar to Rayleigh distribution. The rate of each relative position is higher for happiness.

Achieved results for other vowels are more a less similar to results obtained for /a/ vowel with significant differences of histogram shapes within the emotional state database.

4. CONCLUSION

In this paper were presented the method for recognition the actual emotional state of speaker based on relative position of changes of important points in LPC spectrum. The significant difference between the histogram shape and also its parameters within the emotional states database was proved. Best results have been achieved in the frequency band generated by formants F1 and F2 because these two formants determine the spoken vowel and they are the most common. The higher formants determine the speaker identity and it is possible the higher formant is not generated in the addiction on the actual emotional state of speaker. For all vowels has been investigated the more a less same knowledge.

The normal speaker mood is typical by high rate of each relative position and narrow shape of created histograms. These features have been obtained for all investigated emotional states (anger, sadness, happiness and alcohol intoxication). Within extracted histograms for all states exist significant differences. On the other hand the biggest similarity is between happiness and alcohol intoxication but both states are significantly different. All analysed emotions have been performed by one the voluntary theatre actor.

For future work is necessary to expand the database by higher amount of speakers and real emotional states. All obtained results will be statistically analysed for achievement of common values. The uniformity of common values will be tested by statistical methods F-ratio and p-value which can give the base of further decision toll for emotional state recognition.

ACKNOWLEDGEMENT

The research was performed in laboratories supported by the WICOMT; the registration number CZ.1.07/2.3.00/20.0007, financed by the operational program Education for Competitiveness.

REFERENCES

[1] Hyun, K.: Emotional Feature Extraction Based On Phoneme Information for Speech Emotion Recognition. In: The 16th IEEE International Symposium on Robot and Human interactive Communication, Daejeon, Korea, 2007, s. 802-806

[2] Bozkurt, E.: Formant position based weighted spectral features for emotion recognition. Speech Communication, Volume 53, Issues 9-10, 2011, s. 1186-1197, ISSN 0167-6393

[3] Wang, Y.: Emotional feature analysis and recognition in multilingual speech signal. In: 9th International Conference on Electronic Measurement & Instruments, Jinan, China, 2009, s. 4- 1046-4- 1050

[4] Bogdan, V.: Vowels formants analysis allows straightforward detection of high arousal emotions. In: 2011 IEEE International Conference on Multimedia and Expo, Magdeburg, Germany, 2011, s. 1-6

[5] Razak, A.: A Preliminary speech analysis for recognizing emotion. In: Proceedings. Student Conference on Research and Development, Malaysia, 2003, s. 49-54

[6] Pan, Y. et al.: Emotion-detecting Based Model Selection for Emotional Speech Recognition. In: IMACS Multiconference on Computational Engineering in Systems Applications, Beijing, China, 2006, s. 2169-2172

[7] Psutka, J.: Mluvíme s počítačem česky, Praha, Academia 2006, ISBN 80-200-1309-1