4

Determination of Spectral Parameters of Speech Signal by Goertzel Algorithm

Božo Tomas[1,2,3] and Darko Zelenika[2,3]
[1]*Croatian Telecom d.d. Mostar*
[2]*University "Herzegovina", Faculty of Social Sciences Dr. Milenko Brkić*
[3]*University of Mostar, Faculty of Mechanical Engineering and Computing*
Bosnia and Herzegovina

1. Introduction

The speech is a sound different from all other sounds, and information transmission by speech is a basic mechanism of human communication. The study of speech as one of the main factors of human communication is a multidisciplinary problem thus different fields of science deal with particular aspects of this phenomenon. At the beginning of 21st century the biggest scientific challenge in the segment of speech technologies is the realization of spontaneous communication between human and computer. Technology's solution of recognition and synthesis of speech can't compare with human perceptions and production of the speech. That means our nature, God given, emotion and acts aren't still enough explored, and we have here a lot of space for learning, watching and scientific exploring.

With talking we can transmit to our surroundings complicated information expressed by linguistic contents, but that aren't the least level communication's possibility of speech sound. Sound, which transfers speech signal, carries many information which speech signal contains and decodes that information as well as their production is spontaneous and very simple for human perception (Tomas, 2006). Furthermore, identification, evaluation and selection of certain information contents from a total audio picture are at a consciousness level. Besides linguistic information, the speech signal contains a variety of non-linguistic information from which a listener can get a great deal of information that is not contained in the linguistic information, such as: gender, speaker's age, intentions, a psychological state, and situation the speaker is in; an emotional state of the speaker; surrounding the speaker is in, etc (Tomas et al., 2007a; 2007b; 2009).

Emotions are definitely one of the most important non-linguistic speech attributes (Dellaert & Waibel, 1996). Speaker's emotional state has influence on the articulation and phonation pronounced phonemes. A vocal fold vibration and articulators movement depends on emotional state of speaker. Each speaker's emotion forms vocal chords and vocal system that is acoustically shown through variations of the speech signal parameters. The main problem is to define the parameters that create certain illusions in a speech signal, namely to define correlations between emotions and measurable variations of speech signal parameters (Amir & Ron, 1998; Ramamohan & Dandapt, 2006; Tao et al., 2006). In last few years, numerous investigations have been done in order to determine correlation of the speech parameters and the speaker's emotional state (Amir, 2001; Amir & Ron, 1998; Cowie et al., 2001; Lee & Narayanan, 2005; Morrison et al., 2007; Petrushin, 2000; Scherer, 2003; Ser et al., 2008; Ververidis & Kotropoulos, 2006; Wiliams & Stevens, 1972).

In this chapter, the spectral parameters during the pronouncing of vowels /a/ and /e/ in Croatian (Serbian, Bosnian) language in different emotional conditions has been analyzed. This is done by analyzing structure of vowels /a/ and /e/ spectral parameters for three emotional states: neutral, anger and speech under stress. Impact of emotions on speech signals parameters was analyzed: formant frequencies structure, pitch frequency, dynamics of pitch parameters and pitch harmonics structure. This chapter also describes the Goertzel algorithm and its implementation with speech signal spectral analysis. It describes the method of formant analysis of vowels as well as the vowel harmonics analysis of Goertzel algorithm. The Goertzel algorithm provides fast and simple determination of speech signal structure and precisely determines speech signal frequency values independently of speaker. In analysis of certain speech signal parameters, non-linguistic speech signal attributes can be recognized. The purpose of introducing Goertzel algorithm into the frequency analysis of the digital speech signal is a fast signal processing, and determining of spectral energy from the digital signal in desired frequency bins on the basis of pre-defined coefficient m.

The chapter is organized as follows. In Section 2 the phonetic characteristics of the Croatian language are shortly presented. In Section 3 recording of speech sound materials as well as Goertzel algorithm are shortly presented. Spectral energy analysis of isolated vowels and formant vowel structure tracking by Goertzel algorithm is presented in Section 4. In Section 5, the glottal speech during the pronouncing vowel /a/ has been analyzed. The correlation of certain pitch harmonics parameters and speaker's emotional states were investigated. In Section 6, an imaginary idealized futurist communication model of speech transmission is presented. Section 7 concludes the chapter with final remarks.

2. The basic phonetic properties of the Croatian language

The knowledge of linguistics and phonetics is of great importance in a large number of applications of digital speech processing, such as synthesis and speech recognition (Delic et al., 2010; Flanagan, 1972; Ipsic & Martincic-Ipsic, 2010 & Pekar et al., 2010). The study of linguistics deals with language rules and their impact on human communication, while phonetics deals with the study and classification of sounds in speech.

Language is the basic human communication and it can be transferred by speech or writing-reading. The smallest segment of speech sound is the phoneme. The voice (phoneme) in written language is a letter (grapheme). The voice is the smallest spoken unit that can be isolated from a word of some language, i.e. the voice represents the smallest noticeable discrete segment of sound in continuous speech flow. It is defined as articulated sound in speech which represents the material realization of an abstract linguistic unit - phoneme.

The phonemes are language-specific units and thus each language needs a declaration of its own phonetic alphabet (Sigmund, 2009). The number of phonemes commonly in use in each literary language varies between 30 and 50.

The number of phonemes and graphemes (letter) in Croatian language is the same (30) and they are in alphabetic order: a, b, c, č, ć, d, dž, đ, e, f, g, h, i, j, k, l, lj, m, n, nj, o, p, r, s, š, t, u, v, z, ž. In Croatian as well as kindred South Slavic language (Serbian, Bosnian) any word can be graphically represented using 30 letters (grapheme). Acoustic realization i.e. pronunciation of graphemes is very simple in Croatian language. Each written symbol (letter i.e. grapheme) has its own phoneme and vice versa. This rule stands for all words in Croatian language. Therefore reading and writing in Croatian language is very simple.

Based on the configuration and the opening of the vocal tract, Croatian sounds are divided into three basic groups. These are:

- vowels ('open' sounds): a, e, i, o, u
- semi vowels (sonants): j, l, lj, m, n, nj, r, v
- consonants: b, c, č, ć, d, dž, đ, f, g, h, k, p, s, š, t, z, ž

Semi vowels and consonants when used less specifically are put into the same group, i.e., the consonants.

There are several classifications of the sounds in Croatian language. There are voiced speech sounds and unvoiced speech sounds. Voiced speech sounds classification is the most important in speech technologies. In speech, pitch is not present in all sounds (Ahmadi & McLoughlin, 2010). Pitch harmonics are of a very high intensity in vowels, less high intensity in semi vowels and the least intensity in voiced consonants (b, d, dž, đ, g, z, ž). The rest of the consonants (c, č, ć, f, h, k, p, s, š, t) are unvoiced speech sounds and they have no pitch.

In spoken language the vowels are sound that carry the most information. Vowels constitute the most important speech sounds group and are characterized by the fact that these are the sounds of the greatest energy. When the vowels are produced the majority of the vocal tract is open and in the course of the entire duration of pronunciation the vocal cords vibrate. The primary purpose of vowels is to connect the consonants into the syllables, i.e., the formation of utterable words. In written language, the records of vowels often carry little information, i.e., in most cases the recognition of text messages is possible even when they are completely eliminated out of words.

2.1 The remaining phonetic properties of the Croatian language

Croatian orthographic rules are based on the phonological-morphological principle which enables automatisation of phonetic transcription. Standard definition of orthographic to phonetic rules, one grapheme to one phonetic symbol (Ipsic & Martincic-Ipsic, 2010). Although there are only 30 different sounds in the language, a far greater number of modifications of the same appear in a real speech. The manner of articulation of each sound depends significantly on its context, i.e., the sounds on its left and right side. This phenomenon is called coarticulation. Therefore, high-quality synthetic speech cannot be obtained by simply merging the 30 discreet sounds. Also it is important to emphasize that the transitions from one sound to another are not sudden (step), but are very gradual and are defined by the gradual transition of the articulator from the initial position corresponding the first sound towards the new position corresponding to the next sound. In this process, the vocal tract passes a series of inter-states, which causes the formation of a series of transitive sounds of relatively short duration. The elimination of these transitions significantly disrupts the naturalness of a synthetic speech.

For the purpose of solving this problem, and with the simpler speech synthesizers, the removed pairs of phonemes or so called diphthongs recorded from the actual speech are used as the basic elements of synthesis. In this way, among the basic elements, there are present and also the transitions mentioned above.

2.2 The segmenting of sounds in a continuous natural speech

The uttering of vowels takes from 50 to 300 ms. Consonants are only the processes of in-vibrating and out-vibrating of the previous and the following vowels, with a duration of 2 to

40 ms. The shortest speech sounds last only as it's necessary for the ear to recognize the tonal pitch. The sounds in speech and music should be long enough for the ear to analyze them tonally, about at the same time they must not follow each other too fast so as to prevent masking of the next by the previous masked.

For the research of the variations of the duration of sounds in speech segments during the expression of speech emotions, it is indispensable to analyze the uttered segments on the words as the units of the linguistic context as well as segment sounds from the isolated uttered words. Since there have been noted problems of dichotomy between spoken words and words as units of context in the speech continuum, together with the problems of phonetic positions, the communication situation, individual variations and the like, it is important for the needs of analysis of the manifestations of speech expressions of emotions, though preliminary, to establish criteria for the segmentation of sounds in a continuous speech.

3. Analysis procedure

3.1 Speech material

Twenty students were chosen for this research (all male). The acoustic recordings were made in speech recording and processing Croatian Telecom Mostar studio on mixing board (16-channel MIC/LINE mixer Mackie 1604-VLZ PRO). Each student was asked to pronounce five speech phrases. They are four words: "mama", "ma", "je", "ne" and loudly pronounced vowel "a". A Croatian word "mama" consists of two equal syllables "ma" and it means mother i.e. mum. Word "ma" is a mono syllable word and it is often used to express anger. Word "je" means OK and word "ne" means NO. A certain emotional state was simulated for each speech phrase. The word "mama" was pronounced in a neutrally emotional mode with a bit of sadness, while the word "ma" was pronounced simulating anger and surprise. Also, the pronunciation of the word "je" was without emotions i.e. neutral voice while in pronouncing the word "ne" anger was simulated. Loudness as the expression of stress was simulated by loudly pronounced vowel /a/ in the ˝Jako A˝ speech file. Recording of vowel /a/ in the ˝Jako A˝ speech file was realized in different conditions due to dynamic of the loudly pronounced vowel (less sensitive microphone). The stress conditions considered in this study include simulated anger and loudness (Bou-Ghazale & Hansen, 2000).

The best speaker was chosen by audio testing. Each speech phrase of all 20 speakers (students) was assessed on a scale from 1 to 5. Points were determined for each speaker by taking the sum of the marks for each of his/her five speech phrases. The speaker with the highest score needed to record given speech material another four times. (This consisted of 20 files, 5 for each speech phrase) Finally the best recording of each speech phrase was chosen for a sample. Speech files ˝Jako A˝, ˝MAMA˝, ˝MA˝, "JE" and "NE", were being recorded in Sound Recorder program with sampling frequency of f_u=8000 Hz and 16 bits of mono-configuration and resolution of quantization. Further, the vowel /a/ was isolated from the MAMA and MA files. That is the first /a/ in the word MAMA. In the same way vowel /e/ was isolated from the files JE and NE.

3.2 Spectral analysis

Each temporal signal carries certain frequency contents. The presentation of temporal signals in the frequency domain is very significant. Frequency representation of a signal mostly enables better analysis of appearance which the signal represents. The speech is signal with time-dependent spectral content. Time–frequency representations are often used for the analysis of speech signals due to their non-stationary nature. For a practical application the speech signal can be processed in various ways, other than time- domain, to extract useful information. A classical tool is the Fourier transform (FT) which offers perfect spectral resolution of a signal (Shafi et al., 2009).

Fourier techniques have been a popular analytical tool in the study of physics and engineering for more than two centuries and it is the prime method used to transfer a temporal signal into frequency domain. With the arrival of digital computers, it became theoretically possible to calculate the Fourier series and Fourier transform of a function numerically (Dutt, 1991). A major break-through in overcoming this difficulty was the development of the Fast Fourier Transform (FFT) algorithm in the 1960s which established Fourier analysis as a useful and practical numerical tool. The FFT converts a time-domain sequence x(n) into an equivalent sequence X(k) in the frequency- domain. Spectral analysis of the speech signal in most of published studies was obtained by applying a Fast Fourier Transform (FFT). In this study Goertzel algorithm is used in speech signal spectral analysis.

3.2.1 The application of Goertzel algorithm at the analysis of speech signal

Goertzel algorithm is often applied technique at the realization of digital DTMF (Dual –Tone Multiple Frequency) receivers. DTMF signalling, the so-called tone dialling, is done running audio tones that transmitter generates and that needs to be decoded on the receiver’s side for the purpose of the further processing. DTMF transmitter (encoder) generates a compound audio signal formed of two mutually harmonically independent frequencies by combining of eight given frequencies. Receiver needs to decode the frequency contents of the received audio signal.

DFT plays an important role in the implementation of algorithms into the systems for digital signal processing. The usage of the FFT algorithm significantly reduces the computation time of Fourier transformation. If you need to detect one or more tones in the audio signal or only one or a few frequencies, there is a lot faster method. The Goertzel's algorithm allows decoding of a tone (frequency) with much less processor load compared to the Fast Fourier Transform (FFT).

Since the speech is also a complex audio signal, we draw the same conclusions, which means that at the processing of digital voice signal, we can apply the same technologies and algorithms that we will, of course, adapt to the needs of the analysis of speech signal. It is evident that for the analysis of digital voice signal Goertzel's algorithm can be used. The main advantage of this algorithm is that the coefficients in the equation for a particular frequency are fixed, what makes the calculation much simpler. At the analysis of speech we use the algorithm that at the exit, instead of decoding of levels at the desired frequencies, gives the spectral energy of the analyzed frequency band, i.e. Goertzel's algorithm is used for filtering. When determining spectral energy of digital signal it is necessary to remember only two previously calculated values, from the last step N, difference equations of Goertzel's algorithm, which we calculate by the ‘step by step’ method, while for the detection is required to determine the signal spectrum at the end of filtering.

The purpose of introducing Goertzel algorithm into the frequency analysis of the digital voice signal is a fast signal processing, and determining of spectral energy from the digital signal in desired frequency bins on the basis of pre-defined coefficients m. Then, by the analysis of the calculated energies, for particular bins, we can get a lot of information contained in the speech signal. Also, since using the Goertzel's algorithm we can determine the parameters required for the recognition of speech, the speaker, emotions and other non-linguistic attributes, it is logical that Goertzel’s algorithm can be implemented for the realization of these activities.

3.2.2 Basic Goertzel algorithm

The basic Goertzel transform was derived from the discrete Fourier transform (DFT). Algorithm was introduced by Gerald Goertzel (1920-2002) in 1958 (Goertzel, 1958). It’s an extremely efficient method of detecting a single frequency component in a block of input data. Figure 1 depicts the signal flow for the basic Goertzel algorithm as each sample is processed (Kiser, 2005).

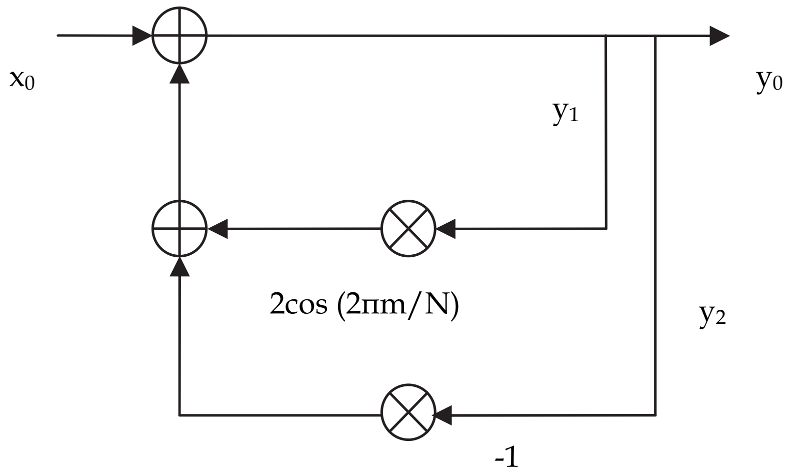


Fig. 1. Signal flow of the Goertzel algorithm

The signal flow of the algorithm produces an output y_0 for each sample processed. The output is a combination of the current ADC sample added to the product of the previous output y_1 multiplied by a constant minus the previous output y_2. Figure 1 may be written as:

$$y_0 = x_0 + y_1 \times 2\cos\left(\frac{2\pi m}{N}\right) - y_2 \tag{1}$$

y_0 is the current processed output, x_0 is the current ADC sample, y_1 is the previously output, and y_2 is the next previously processed output, m is the frequency domain bin number. N is the sample block size. Input samples are processed on a sample-by-sample basic. Processing continues over a block of input data length N.

The Goertzel transform is normally executed at a fixed sample rate and a fixed value of N. To detect multiple frequencies (i*.e. frequency bins),* each frequency must be assigned its own coefficient and then the value is used in equation (1). Depending on the number of desired detection frequencies, there will be an equal number of equations like equation (1) that must be executed once during each sample. At the end of a block of data, the spectral energy of each frequency bin will be computed and validated.

3.2.3 Spectral energy of Goertzel bins

The Goertzel algorithm is a filter bank implementation that directly calculates one Discrete Fourier transform (DFT) coefficient. The Goertzel algorithm is a second-order filter that extracts the energy present at a specific frequency (Bagchi & Mitra, 1995; Felder et al., 1998). Therefore, for the analysis of digital speech signal a system that will give a spectral energy of the analyzed (filtered) signal bandwidth (bin) is suitable. Such system is shown in Figure 2.

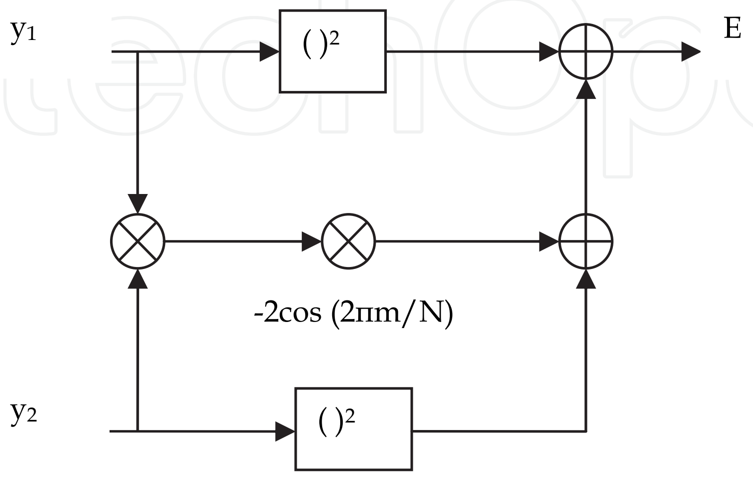


Fig. 2. Determination of spectral energy bin

Figure 2 may be written as:

$$E_m = y_1^2 + y_2^2 - 2y_1y_2\cos\left(\frac{2\pi m}{N}\right) \tag{2}$$

After a block of data has been processed, spectral energy for the signal of interest (frequency bin) is determined by y_1 and y_2 variables. The sum of the squares of y_1 and y_2 are computed to determine the spectral energy of a particular frequency bin. One of the advantages of the Goertzel transform is that spectral energy computation needs to be performed only at the end of a block of data. After a block of data has passed through equation (1) (filter part), the spectral energy of Goertzel bins (frequency bin) is simply determined by equation (2) (energy part).

The Goertzel algorithm implementation has the tremendous advantage that it can process the input data as it arrives. The output value is only needed for the last sample in the block. The FFT has to wait until the entire sample block has arrived. Therefore, the Goertzel algorithm reduces the data memory required significantly.

In the analyses, spectral energy distribution of isolated vowel /a/ is in the final invariant time position during its pronunciation. Our software built for these researches allows selecting of sample, from the analyzed file, from which it will start loading samples in the equation (1). The analyses (all five speech files of chosen speaker) begin from sample k=50 which mean that the first 6,25 ms of speech file are skipped. Thereby, beginning transition of vowel pronunciation was avoided. The speech file is first loaded into the software, after that frequencies tuning coefficients – *m* are loaded, followed by sensing start sample – *k* and number of samples – *N* that is number of Goertzel transform iteration i.e. the equation (1).

4. Formant vowel structure tracking by Goertzel algorithm

In many science papers from this area, for the most significant parameters of speech signal which fertilized change under emotions and stress there has been defined basic frequency and energetic structure of pronounced words. Also, and analysis of formants show that exist considerable departure in formants structure at pronounced speech in different emotion’s state or under impact of stress. Spectral energy analysis of isolated vowels by applying Goertzel’s algorithm gives a proper illustration that enables classification of emotions by comparing the energy, and their formants.

Formant frequencies structure gives a good view to determine linguistic as well as non-linguistic speech meaning and it represents a significant parameter within the synthesis and speech recognition procedure. In fact formants are groups of harmonics enhanced by the vocal tract resonance. There is a resonance i.e. the enhancement of spectral energy in speech signal spectrum at formant frequencies (Tomas & Obad, 2009).

The difference between formant frequency structures of vowel /a/ is very informative about the speaker’s emotional state. The dominant influence of emotions is on the first formant. Frequencies of the first formant are increased in expression of the emotions of anger and happiness but decreased in case of fear and sadness. Therefore, changes of vowel /a/ formant structure are not sufficient for separation of primary emotions: anger, happiness, fear and sadness.

4.1 The determination of the formant vowels structure using the Goertzel algorithm

We will analyze the speech signal, vowel / a /, recorded with the “Sound Recorder” software with the sampling frequency fu = 8000Hz, and stored under the title "Jako A". If the Goertzel's transformation, equation (1), is performed on the time segment of the recorded speech signal of 25ms duration, we have N = 200 samples of speech signal, which we process by the algorithms described by the equation (1) and the equation (2). From f_u = 8000Hz and N = 200 result’s the frequency band for each Goertzel's bin: B = f_u / N = 40 Hz. The first formant will be sought in the frequency band (400 - 720) Hz. This frequency band is divided in the next 8 bins. The frequency bands of the eight bins, as well as their central frequencies, on which we base the calculation of the coefficients of tuning the frequency domain m, are shown in Table 1.

Bin	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
B (Hz)	400-440	440-480	480-520	520-560	560-600	600-640	640-680	680-720
f_{bins}	420	460	500	540	580	620	660	700

Table 1. The division of frequency band into bins

Now we will calculate the coefficients of tuning the frequency domain m for each bin, which will be included in the expression 2cos(2πm/N) and the value of that expression, that is a constant - different for each bin, will be stored on particular memory locations, so that the processor spends less time for the calculation of the equation (1), since the equation is executed in 200 steps. We will illustrate this for bin 1:

$$f_{bins}^{(1)} = 420 \Rightarrow m_1 = f_{bins}^{(1)} \frac{N}{f_u} = f_{bins}^{(1)} \frac{200}{8000} = \frac{f_{bins}^{(1)}}{40} = \frac{420}{40} = 10,5 \qquad (3)$$

The coefficient of tuning the frequency domain, the k-th bin m_k, with the sampling frequency f_u = 8000Hz, and the length of block of time samples N = 200, we calculate by the following equation:

$$f_{bins}^{(k)} = define \Rightarrow m_k = f_{bins}^{(k)} \frac{N}{f_u} = f_{bins}^{(k)} \frac{200}{8000} = \frac{f_{bins}^{(k)}}{40} \qquad (4)$$

By the inclusion of the obtained values m_k in the expression of $2\cos(2\pi m_k/N)$, we get for each bin its coefficient $k_{m(k)}$ for the calculation of the equation (1). The outline of all values is given in Table 2 (Tomas & Obad, 2008).

Now we have all the necessary input constants for the calculation of equations (1) and (2). The samples of speech signals are entered in the same order that they are sampled in the equation (1), sample by sample, i.e. the equation is performed step by step until the last step N = 200. The calculated values of the equation in the last two steps N = 200 and N = 199 are stored and entered into the equation (2). For each bin we calculate the spectral energy by entering the coefficients of the bin into the equation (1). In fact, in this way we determine the spectral energy on the frequency of f_{bin} in its corresponding Goertzel's bin.

Bin	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
f_{bins}	420	460	500	540	580	620	660	700
m_k	10,5	11,5	12,5	13,5	14,5	15,5	16,5	17,5
$k_{m(k)}$	1,8921	1,8477	1,8477	1,8228	1,7960	1,7675	1,7372	1,7052
E	4,3561	4,6030	69,4686	74,4163	48,7836	144,986	**416,102**	58,0926

Table 2. Display of parameters, coefficients and spectral energy of bins

The same procedure is for other formants, which would be searched for in higher frequency bands. Also, we do not have to take bins in a row. Also, by the simple choice of constants we calculate the bins that overlap. In addition, if we need greater selectivity of bins we are able to accomplish that by increasing the number of input samples N with unchanged sampling frequency, or vice versa.

From Table 2 it can be seen that the first formant of the vocal /a/ is positioned in the bin 7 in the frequency band (640-680) Hz. Also, based on the value of energies in bins 6 and 8 we read that the formant is positioned at the beginning of the bin 7, i.e. that it is in the frequency band (640-660) Hz. However, for the majority of analysis it is not necessary to determine the exact frequency positions of formants but the division of the bins is quite enough. The conclusions about the emotional states of speakers, as well as the recognition of speech and speaker can be drawn on the basis of relations between the energies of selected bins.

For more precise determination of the frequency of the first formant we can carry out the following procedure of Goertzel's transformation on bins 6 and 7. To do that we would increase the number of samples of speech signal (e.g. N=400) for that frequency band so that the frequency band for new Goertzel's bin would be: B= $f_u/N=8000/400=20$ Hz. The analyzed frequency band (620-680) Hz can be divided into three new bins and for these three new bins we can calculate spectral energies. However, with N = 400, i.e. 50ms of speech signal we would come out of the frame of quasistationariness. On the other hand we could with N = 200 by selecting m-coefficient of tuning of a segment of frequency domain, analyse the bands with overlapping of neighbouring bins, e.g. 50% and get the spectral energy on the analysed frequency band with the distance of 20 Hz instead of 40 Hz as in Table 2 where the bins are in a row with no overlapping.

4.2 Influence of emotions to the formant structure of vowel /a/

Using the results of the (Vojnović, 2004) for Serbian language with the PRAAT program for male speakers, we shall make an analysis applying the Goertzel algorithm. Achieved results of the influence of emotional state of males to the formant structure of vowel /a/ are displayed in the Table 3. A similar analysis was done for German language (Kienast & Sendlmeier, 2000). Hence, with male speakers, no matter the emotional state of a speaker, while pronouncing vowel /a/ formants should be looked for in the following frequency bandwidths:

- Formant 1 in the bandwidth of (400-700) Hz
- Formant 2 in the bandwidth of (1350-1500) Hz
- Formant 3 in the bandwidth of (2500-2700) Hz

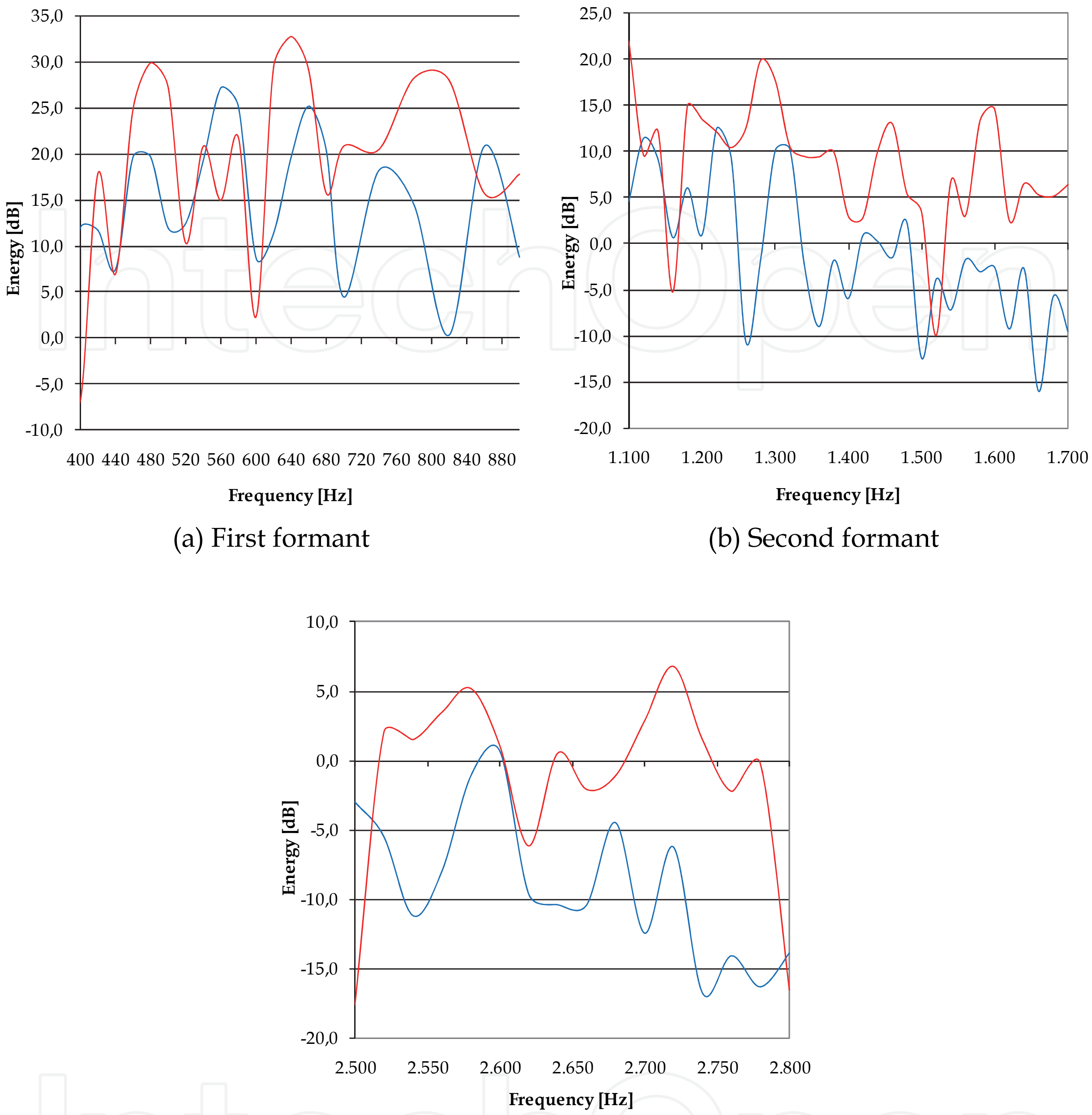
EMOTIONS	Frequency (Hz) of 1st formant	Frequency (Hz) of 2nd formant	Frequency (Hz) of 3rd formant
N -Neutral	539	1393	2560
A- Anger	651	1436	2604
H- Happiness	605	1458	2600
F- Fear	538	1466	2599
S- Sadness	455	1422	2672

Table 3. Average middle frequency of the first three formants of male speakers for different emotional states

Figure 3 illustrate the comparison of spectral energies of the vowel /a/ within expected ranges of the first three formants (the red line is isolated vowel /a/ from "MA" speech file, and the blue line is isolated vowel /a/ from "MAMA" speech file). In comparison of energy values for emotions classifications, in equations (1) and (2) we adjust the bins of wider frequency bandwidth. At the sampling frequency of 8 kHz it is convenient to use N=100 or N=200 and so get the bins in bandwidth of 80Hz and 40 Hz respectively.

4.3 Influence of emotions to the frequency composition of vowels /a/ and /e/

Figure 4 and Figure 5 illustrate the comparison of spectral energies of isolated vowels /a/ and /e/ in neutral pronouncing (MAMA and JE) and in simulated anger (MA and NE) at the bandwidth of Goertzel's bins of 80 Hz (N=100) and 40 Hz (N=200) respectively. While



(a) First formant

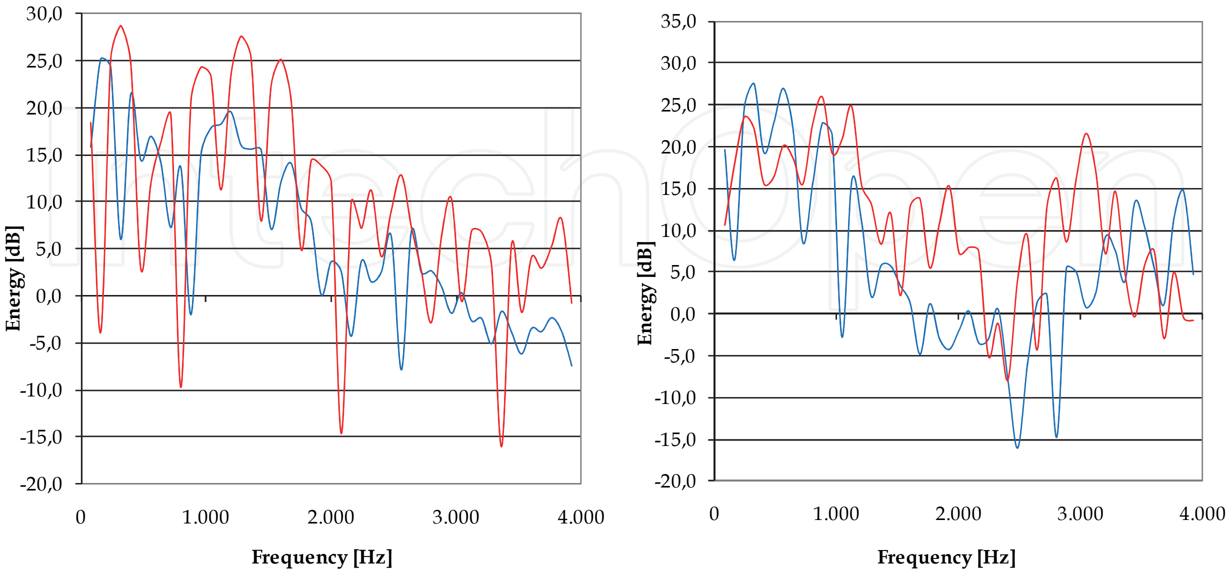
(b) Second formant

(c) Third formant

Fig. 3. Bandwidth of the first three formants "A" N=200, f=8 kHz, overlapping of bins 50%

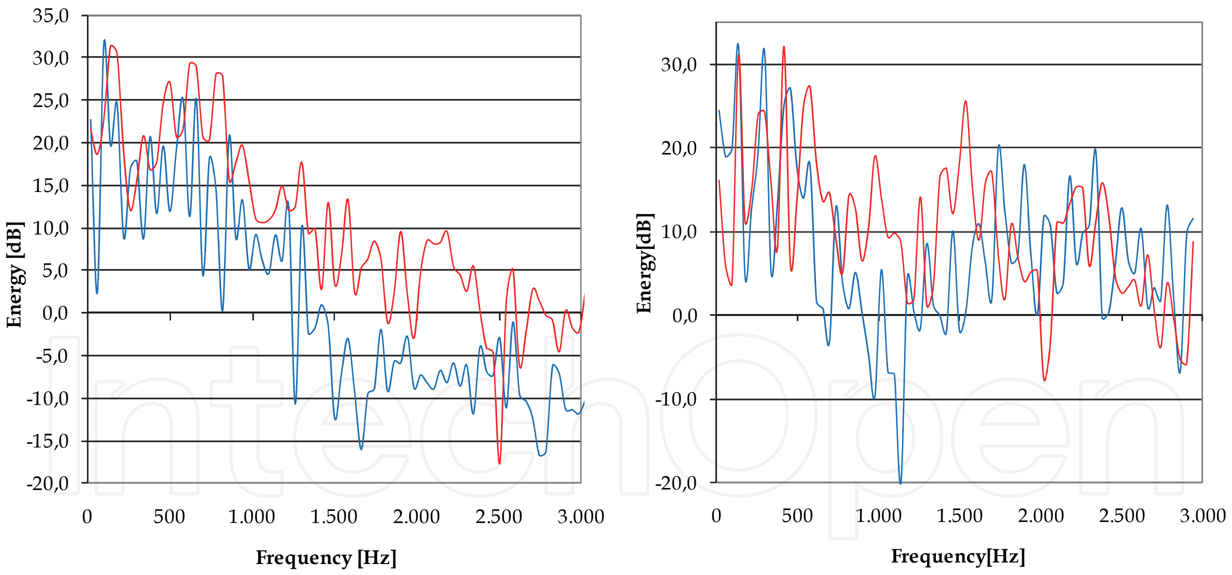
pronouncing /a/ from MA in anger, positions of formants frequency are higher than while pronouncing it without emotions in articulation of word MAMA, which affirms the researches illustrated by the Table 3. Also, it is obvious that the energy of spectrum /a/ from MA, of articulation in anger, is higher that the energy of the neutral articulation of vowel /a/ from MAMA within all three frequency bandwidths observed. Also, as we positioned the first formant for "JakoA" in the bandwidth of 640-660 Hz, we may conclude that "JakoA" has been articulated in the emotional state that might be anger, happiness or some stressful emotion, since F1> 600 Hz (Tomas & Obad, 2009).

On the following figures the blue line is neutral articulation (vowel /a/ is isolated from "MAMA" speech file and vowel /e/ is isolated from "JE" speech file), the red line is articulation that expresses anger (vowel /a/ is isolated from "MA" speech file and vowel /e/ is isolated from "NE" speech file).



(a) Comparison of two articulations of /a/ (b) Comparison of two articulations of /e/

Fig. 4. Comparison of spectral energies of isolated vowels /a/ and /e/, N=100



(a) Comparison of the vowel /a/ N=200 (b) Comparison of the vowel /e/ N=200

Fig. 5. Comparison of the spectral energy of isolated /a/ and /e/, N=200

Generally speaking, with spectral analysis of isolated vowels /a/ and /e/ from words: MA, MAMA , JE and NE by applying Goertzel's transformation, we may recognize the emotions of the speaker by comparing following parameters of voice signals received spectrums (Tomas & Obad, 2009):

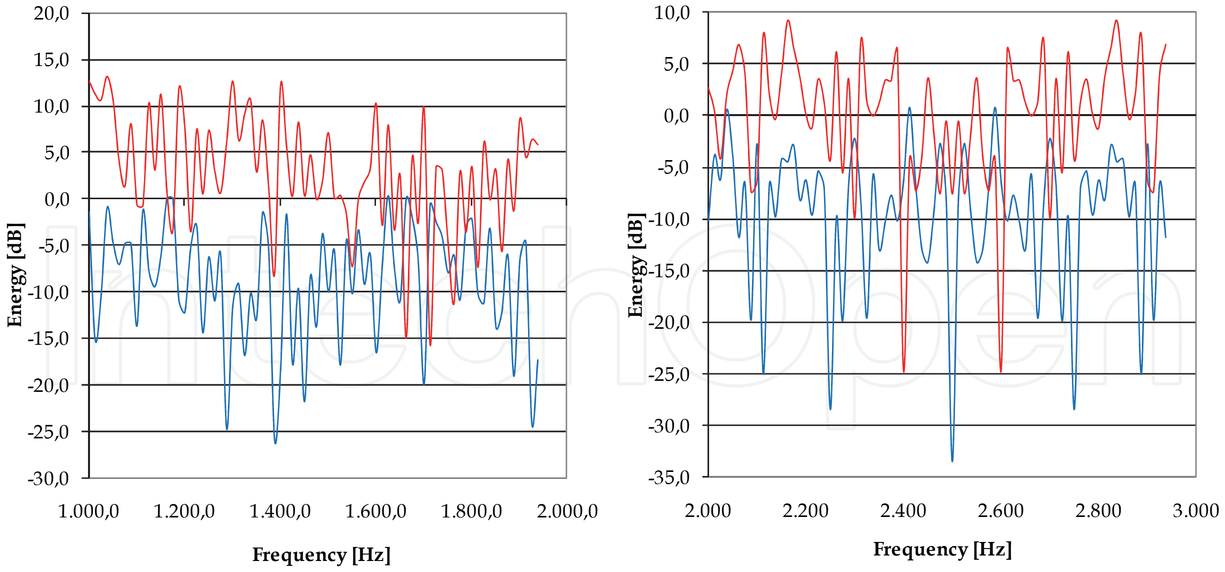
- in articulation without emotions and with sadness and sorrow, spectral energy is mostly lower (in almost all bandwidths) than in articulation with expression of emotions (A,H,F) with exception of bandwidths of low frequencies of the first harmonics.
- hence, one of the ways to identify emotions for one speaker at analysis of spectrum of vowel /a/ by Goertzel's algorithm, is comparison of the spectral energy of isolated vowel /a/ in frequency bandwidth (above the first harmonic) of few bins (depending on the bandwidth of bin) with the energy of referent signal of vowel /a/ neutral i.e. without emotions. Here we compare the ratio of sum of energy bins of analyzed /a/ (Ean) with the sum of energy bins of neutral referent /a/ (Er). If Ean > Er we then have the expression of A, H or F, and if the ratio Ean/Er <1 we have then the expression of (**S**- sadness) or (**Sr**-Sorrow). We may achieve the same results with tabular view. The similar conclusions may be applicable for vowel/e/.
- in neutral articulation and with emotions (S and Sr) the frequencies of the first formant of vowel /a/ are lower than in articulation with expression of emotions (A and H). With Goertzel's algorithm we affirm the results achieved by other methods: F1>600 Hz A or H and for F1 <600 Hz, it is the case of S or Sr (for males).
- in case of vowel /e/ the formants frequencies for different emotions are more distanced than it is the case for /a/ , which is shown by Figure 4 and Figure 5.

For comparison of spectral energies and formants in classification of emissions we may use the bins of wider frequency bandwidth. For the case of sampling frequency of 8kHz presented bandwidths of 40 Hz and 80Hz defined by the numbers of samples N=200 and N=100 respectively give a proper graphical overview and good results. Advantage of using bins of wider bandwidth is that we cover the analyzed bandwidth with small number of bins that program may process in one cycle since it uses the same input samples. Of course each bin has different coefficients and final values of spectral energy.

It is obvious from the Figure 4 and Figure 5 that the spectral energy of isolated vowel /a/ from MA in articulation that expressed the anger is much higher compared to /a/ from MAMA in neutral articulation that expresses no emotions. It is noticeable as well that in the ranges of lower frequencies that difference of energy is less expressed, while in the range of middle frequencies in the bandwidth of 1-4 kHz the energy difference is expressed the most. It is illustrated by Figure 6, where there are precisely calculated spectral energies in that frequency bandwidth with distance points of 12,5 Hz (N=320, f= 8 kHz, overlapping of bins 50%).

Figure 6 (a) illustrates the comparison of spectral energies of isolated vowels /a/ digitalized with sampling frequency of 8 kHz, while Goertzel's transformation was applied with 320 samples of digitalized signal so that the bins were of 25 Hz bandwidths.

Except this, the neighbouring bins are overlapping 50% so the spectral energy was calculated for the frequency bandwidth of 1-2 kHz with calculating distance points of 12,5 Hz. Figure 6 (b) illustrates the comparison of spectral energies of isolated vowel /a/ in the bandwidth of 2-3 kHz. Hence, if we perform the identification and classification of emotions with only energy comparisons of isolated vowels /a/, we may do a quick analysis by selecting several bins in string from the frequency bandwidth of middle frequencies (1-4 kHz) and to compare the sums of the energies of those bins as it was explained earlier (Tomas & Obad, 2009).



(a) Bandwidth of 1-2

(b) Bandwidth of 2-3

Fig. 6. Comparison of spectral energies of vowels /a/ in the bandwidth of 1-2 kHz and 2-3 kHz

5. Pitch harmonic parameters

Spectral energy analysis of isolated vowels by applying Goertzel's algorithm enables classification of emotions by comparing their harmonics (Tomas et al., 2007b). For selected number of samples *N=320* duration of time frame is 40 ms, which still satisfies quasistationary speech conditions. Harmonics parameters of vowel /a/ isolated from ˝Jako A˝, ˝MAMA˝ and ˝MA˝ speech files are shown in Table 4.

Emotion s	H1	H 2	H 3	H 4	H 5	$D=H_{max}-H_{min}$	F (Hz)
MAMA	1	3	2	5	4	3332,5-47,8=3184,7	58,75
MA	1	5	4	2	3	7855,7-270,1=7585,6	97,5
Jako A	2	4	5	3	1		80,9

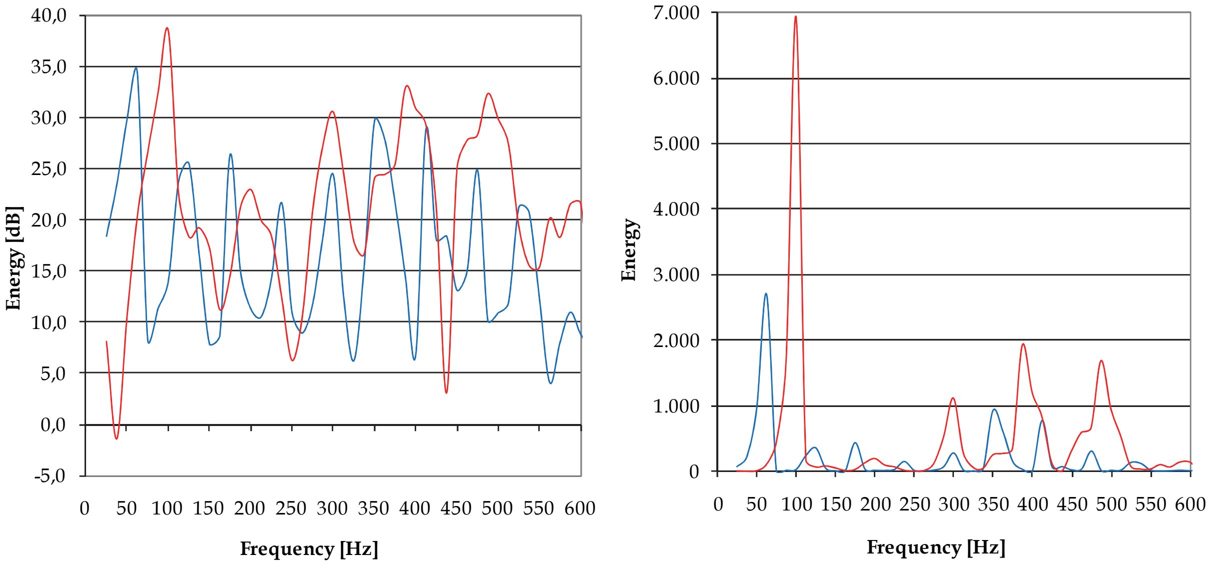
Table 4. Structure, dynamics and basic frequency of vowel /a/ harmonics depend on emotions

Conclusions of the first five harmonics parameters analysis are:

- Harmonics frequencies depend on speaker's emotions
- Distribution of harmonics energy amplitudes (harmonics amplitude structure) depend on emotions (1st harmonic has the highest energy value, 2nd harmonic has the second high energy value, and energy value declines respectively, so 5th harmonic has the lowest energy value.
- Histograms and dynamics of harmonics energy amplitudes depend on emotions.

Recording of vowel /a/ in the ˝Jako A˝ speech file, was realized in different conditions (less sensitive microphone), therefore absolute and decibel measurements values are not identical to values of vowel /a/ in ˝MAMA˝ and ˝MA˝ speech files. Thus, their dynamic and histogram bands are not comparable. However, these changes do not influence frequency and harmonic structure.

The following Figures illustrate vowel /a/ harmonics. There is dB and absolute amount of spectral energy in Figure 7 and Figure 8. Figure 7 shows frequency band 0-600 Hz, N=320 and frequency bin overlap of 50 %. Sampling frequency is 8 kHz, thus bin width is 25 Hz, and spectral energy results are shown by 12,5 Hz points distance. On Figure 7 the blue line is neutral articulation (vowel /a/ is isolated from "MAMA" speech file), the red line is articulation that expresses anger (vowel /a/ is isolated from "MA" speech file).



(a) Display in decibels

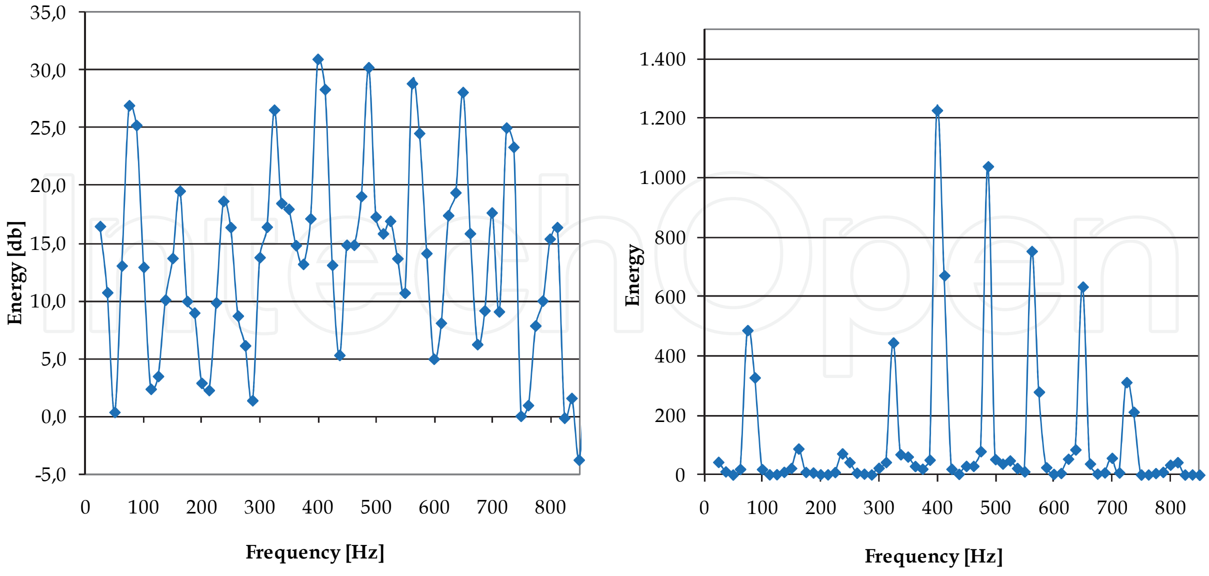
(b) Display in absolute values

Fig. 7. Comparison of vowel /a/ harmonics

Figure 8 shows spectral energy diagram in frequency band of the first ten harmonics of vowel /a/ in the ˝JakoA˝ file. Spectral energy is shown in dB (10 logEaps).

Figure 7 and Figure 8 illustrate correlation of harmonics amplitude, dynamic and base frequency with emotions. In Table 4, the lowest basic frequency has vowel /a/ isolated from the MAMA file, spoken in a neutral emotional condition, while the highest basic frequency has vowel /a/ isolated from the MA file, spoken with angry and surprise simulation. Amplitudes dynamic in angry is higher than in neutral emotional speech. Also, amplitudes dynamic of ˝Jako A˝ file would have been higher too if it had been recorded in the same conditions. Harmonics structure in all three emotional conditions is different indicating emotional influence, but for determination of correlations more analyses are required. It is necessary to import parameters – indexes that will clearly describe harmonics structure.

The Goertzel algorithm provides fast and simple determination of speech signal structure and precisely determination of speech signal frequency values independently of speaker. In analyzes of certain speech signal parameters, non-linguistic speech signal attributes can be recognized (Tomas et al., 2007b).



(a) Display in decibels (b) Display in absolute values

Fig. 8. Harmonics diagram of ˝Jako A˝ file

5.1 Harmonics structure normed indexes (HSNI)

Implementation of parameters provides fast, efficient and clear overview of analyzed speech files harmonics structure. HSNI are parameters that are fast and simply calculated and they provide articulate and fast classification of emotions (Tomas et al., 2007a). Table 4 shows structure of the first five harmonics. Numbers, that defined intercomparison of the absolute mounts of energies, are assigned to harmonics. Energies are identified by Goertzel algorithm with 12,5 Hz distance frequency, at the harmonic frequency. If energies are identified with higher calculating precision of 2,5 Hz, for our speech files amounts are shown in the Tables 5, 6 and 7. In the Tables there are exact amount of frequencies at which spectral energies amount are maximal. It is evident that harmonics structure in Tables 5. and 7. are changed after more precise calculations with harmonics structure in the Table 4. There are permutations of the harmonics H2 and H3 in the Table 3 and H4 and H5 in the Table 7 respectively.

Harmonics	H1	H2	H3	H4	H5
Structure	1	2	3	5	4
Freq. (Hz)	58,75	120	176,25	236,75	297,5
Energy	3332,5	523,4	443,2	147,8	315,4
Index	**1**	**0,157**	**0,132**	**0,044**	**0,095**

Table 5. The first five harmonics structure of the vowel /a/ from ˝MAMA˝ file

Harmonics	H1	H2	H3	H4	H5
Structure	1	5	4	2	3
Freq. (Hz)	97,5	195	297,5	390	487,5
Energy	7855,7	270,1	1172,6	1941,8	1688,4
Index	**1**	**0,0344**	**0,1493**	**0,2472**	**0,2149**

Table 6. The first five harmonics structure of the vowel /a/ from ˝MA˝ file

Harmonics	H1	H2	H3	H4	H5
Structure	2	5	4	3	1
Freq. (Hz)	80,6	161,25	242,5	325	400
Energy	713,3	91,2	113,8	445,6	1227,2
Index	**0,581**	**0,0743**	**0,093**	**0,363**	**1**

Table 7. The first five harmonics structure of the vowel /a/ from ˝ Jako A ˝ file.

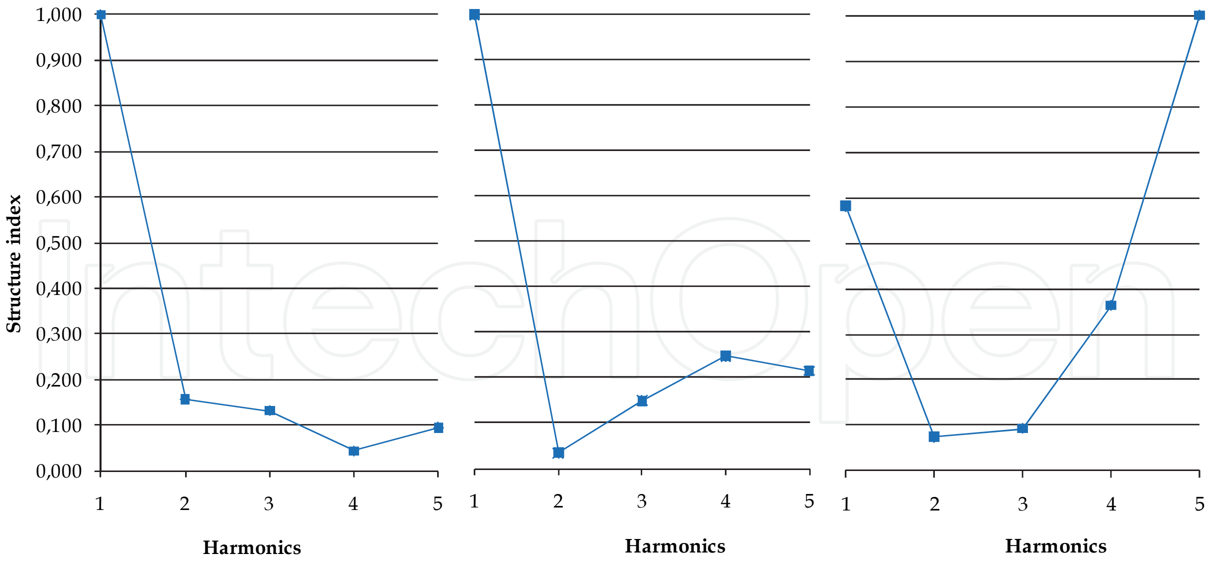
Harmonics	H1	H2	H3	H4	H5
MAMA Index I_k	**1**	**0,157**	**0,132**	**0,044**	**0,095**
MA Index I_k	**1**	**0,0344**	**0,1493**	**0,2472**	**0,2149**
"Jako A"Index I_k	**0,581**	**0,0743**	**0,093**	**0,363**	**1**

Table 8. The first five HSNI of the vowel /a/

Absolute amounts of spectral energies are normed thus spectral energies amounts of every harmonics divide with maximal amount harmonic spectral energy. That defines parameters shown in the last rows of Tables 5., 6. and 7 . These parameters are called HSNI and they are easily defined by equation (5):

$$I_k = \frac{E(H_k)}{E(H_{\max})} \qquad (5)$$

I_k is index of the H_k harmonic, $E(H_k)$ is spectral energy amount of the H_k harmonic and $E(H_{max})$ is spectral energy amount with the highest absolute amount. In the Tables 5, 6 and 7 calculated amount of I_k index are shown in the last rows and amount of $E(H_k)$ are shown in the next to last rows of the Tables. These indexes are parameters for recognition and classification of emotions and others non-linguistic speech attributes (Tomas et al., 2007a). Graphs of HSNI linear interpolation are shown on the Figure 9.



(a) /a/ from "MAMA" file (b) /a/ from "MA" file (c) /a/ from "Jako A" file

Fig. 9. Graphs of HSNI of vowel /a/

HSNI of vowel /a/ from ˝MAMA˝, "MA˝ and ˝Jako A˝ speech files are shown in Table 8. Also, graph of indexes linear interpolation is shown in Figure 10.

The graphs confirmed that implementation and analyze of these indexes provide recognition and classification and others non-linguistic speech attributes.

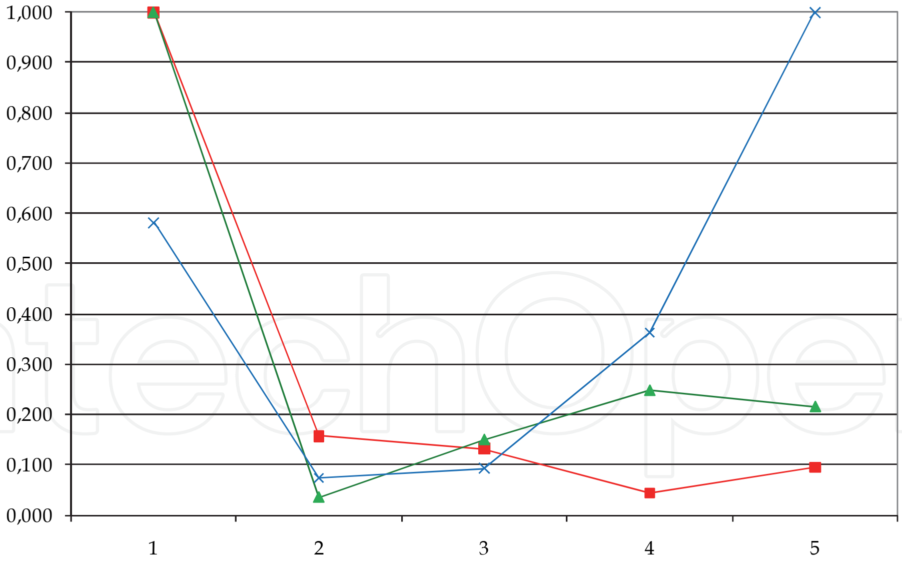
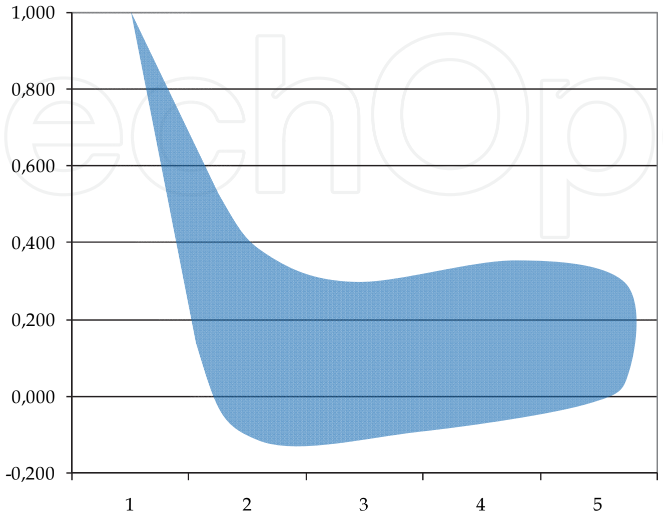


Fig. 10. HSNI of vowel /a/ from ˝MAMA˝, MA˝ and ˝Jako A˝ file

5.2 The area of stress

The lines on figures 9 and 10 show graphs of HSNI of vowel /a/ for our chosen speaker. The graph (lines) of indexes linear interpolation of vowel /a/ is different depending on the speaker's emotional state. One can assume that the graphs of other vowels will behave similarly. That means that the curves of all five vowels, articulated by one speaker, for a certain emotion will almost overlap. Therefore, we will have areas of emotions. The area of stress and the expected areas of articulation without emotions (without stress) for our speaker are shown on the following figure.



(a) Tracking the speaker "without stress" (b) Tracking the speaker "with stress"

Fig. 11. Expected areas of the articulation

If we track a speaker's HSNI graph during continuous speech, (without emotions and stress), the graph will be in the blue area. However, if for any reason the speaker experiences stress (ex. A pilot experiencing stress because of an unexpected situation on flight) the HSNI graph will no longer be in the blue area but rather in the red area which is the area of stress.

In this chapter, HSNI, the lines of emotions, and the area of stress are presented on the basis of the first five harmonics. To get better results, it would be desirable to work with a larger number of harmonics. In order to detect stress in our speaker, only the first and higher (4 and 5) harmonics indexes are relevant, while the second and third indexes stay in the same area whether the speaker experiences stress or not. Therefore, by tracking the shape of HSNI lines and the area in which the lines are located, stress of a speaker can be detected. When tracking the area of stress, instead of tracking the lines we can track HSNI points only. As far as our speaker is concerned we should track when the first, fourth and fifth HSNI point enters the red area. Expected areas with and without stress for our speaker are shown on Figure 12.

In many applications the emotional state of one speaker (the pilot) is tracked. It is especially very important to determine when that speaker undergoes stress (Hansen et al., 2000; Zhou et al., 2001). It is shown how the harmonic frequencies change if the speaker changes emotional states. Besides that, during speech in one emotional state the basic speech frequency is not fixed. Therefore, the HSNI and the areas of emotions are practical for the classification of emotions and especially for stress detection.

It is known that the frequencies of basic harmonics are different for male, female and child speakers. The main feature which can speaker's sex distinguish is fundamental frequency *F*0 with typical values of 110 Hz for male speech and 200 Hz for female speech. The pitch of

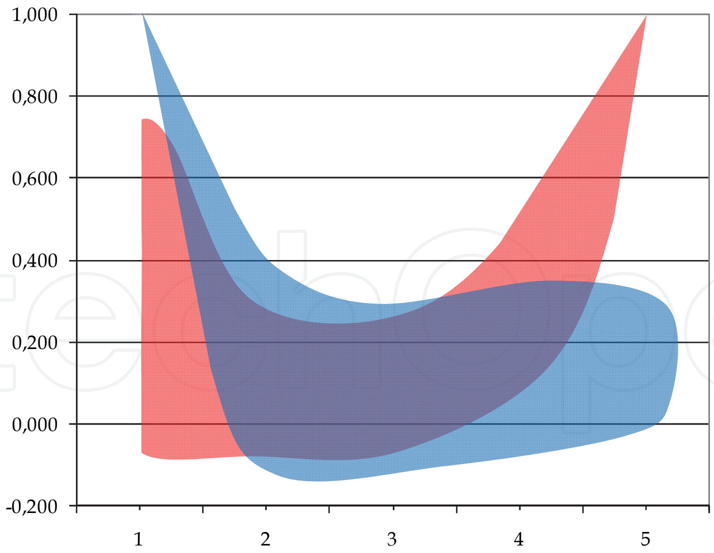


Fig. 12. Expected areas "without stress" and "with stress"

children is so different that they are often treated as "the third sex". Most values of *F*0 among people aged 20 to 70 years lie between 80-170 Hz for men, 150-260 Hz for women and 300-500 Hz for children (Sigmund, 2008). Therefore, in many applications in speech technologies these three groups of speakers must be programmed separately. HSNI can probably help to overcome this problem.

6. The voice communication channel model

Communication with computers, by means of speech has been a topic frequently regarded as a science fiction. In the last thirty years technology that has enabled speech recognition and its synthesis has made exceptional developments. At the same time, we are witnesses of explosion in systems used for recognition and synthesis of speech with an additional dimension: emotions.

If standard voice communication channels are enhanced by compatible ASR and speech synthesisers, which code linguistic and non-linguistic speech parameters with common protocol we get communication model with negligibly short link engagement in comparison to duration of the actual conversation (Tomas et al., 2007a). Transfer of voice information content down the line would take short time, whereas voice through the microphone terminal on the side of speaker and voice through the synthesiser on the side of listener are in real time. Actually, at transfer we use coded indexes of non-linguistic attributes of the speaker's voice (speaker's voice characteristics) and coded indexes of linguistic information content from voice database. These indexes are coded data of linguistic and non-linguistic information with which are passed on the linguistic meaning of the conversation textual content and non-linguistic attributes of the speaker.

This model would have been more practical and simpler in an enclosed system, where the number of participants in communication is limited as well as the pool of words and expressions which are used in communication. Naturally, improvement of systems for synthesis and recognition of voice will enable implementation of this model in open communication systems.

A statement is a pronunciation of one or more words that have a single meaning to the computer. A statement can be one word, several words, a sentence or several sentences. The representation of the described voice communication model is shown in Figures 13 and 14. The Figure 13 shows speaker's side of the model, whereas Figure 14 shows receiving end, i.e. the listener's side.

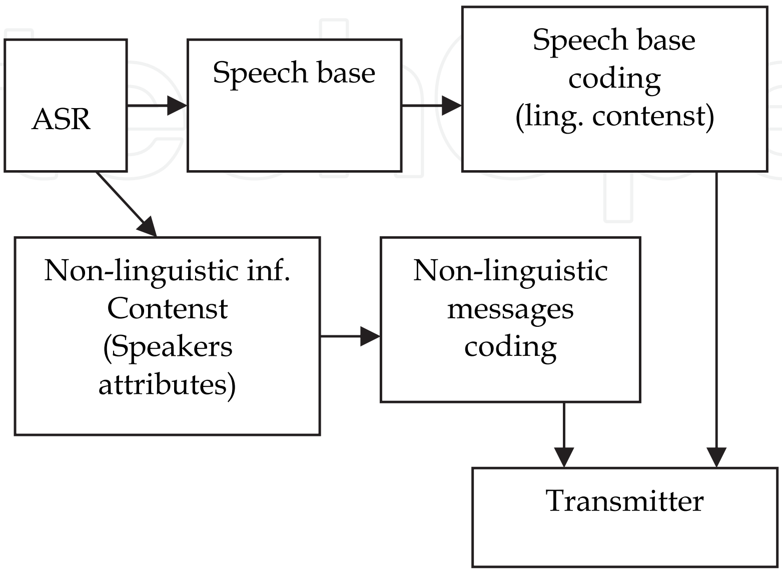

ASR
Speech base
Speech base coding (ling. contenst)
Non-linguistic inf. Contenst (Speakers attributes)
Non-linguistic messages coding
Transmitter

Fig. 13. A part of communication channel, speaker side

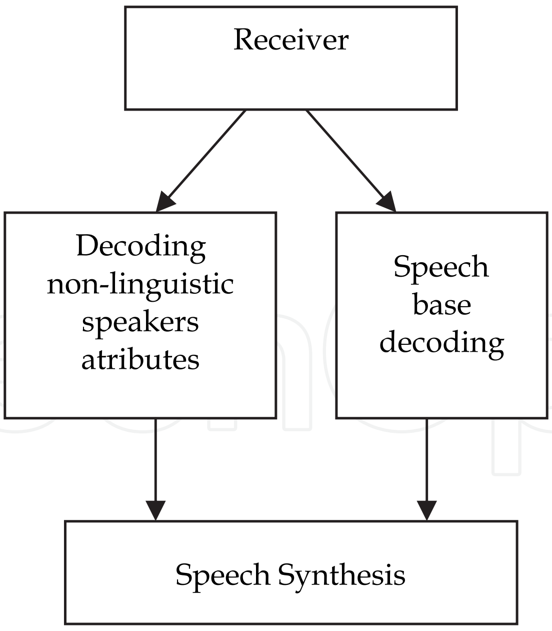


Fig. 14. A part of communication channel, listener side

Let us presume that we have closed communication system with 1000 participants. Every participant have stored and coded relevant non-linguistic characteristics of his voice. After link connection, non-linguistic information coder forwards characteristics to transmitter, which transmits them. On the receive side non-linguistic information decoder recognize speaker and send information about non-linguistic parameters of speakers voice to a synthesizer.

After non-linguistic information exchange, transmitter ASR recognizes spoken words and statements. They are coded in the speech base coder and are forwarded to receiver like coded indexes. On the receiver side, speech base decoder decodes received words and statements. A speech synthesizer produces them with recognized speaker voice. Of course, this basic model can be upgrade depending of communication requirements i.e. (military, police, etc.).

6.1 Model application possibilities

A telephone PSTN channel was used only as an illustration and packet networks would be suitable for speech transmission by this model. It would also be very convenient to divide the information part of the packet into two parts where linguistic speech information could be transmitted through one part and nonlinguistic information could be transmitted through the other, as shown in Figure 15. Due to nature of speech itself, it is not necessary for each packet to contain the nonlinguistic part, so it possible to get even a better efficiency of this model. This is very important with the realization of speech communication through networks with reduced speed and capacity channels such as VoIP that is the most current one today but also the other packet networks and their protocol.

Linguistic Information	Nonlinguistic Information	Header packet

Fig. 15. Packet structure

Application possibilities of this model are multiple and we will presume mention some of them:

- Closed communication system
- Open communication system
- Speech to text
- Conversation of different languages speakers if an interpreter is within the model
- An aid for persons with special needs
- VOiP
- Speech message transformation systems
- Automatic speech translation from Croatian intro other world-wide languages and via versa
- E-mail and SMS messages reading
- Text to speech
- Etc.

Finally, stress and emotions classification can also be employed in forensic speech analysis by law enforcement to assess the state of telephone callers or as an aid in suspect interviews.

The majority of studies in the field of speaker stress and emotion analysis have concentrated on pitch, with several considering spectral features derived from a linear model of speech production.

7. Conclusion

Each speech message, besides linguistic information content (semantic meaning of spoken text), contains non-linguistic characteristics of speaker. The non-linguistic characteristics are correlated to certain acoustic characteristics of speech signal. Studying non-linguistic information's contents in speech signal is quality jump in speech analyses. The aim of many future researches will be finding suitable acoustic characteristics of speech signal that, besides the linguistic contents, will be used for a reliable determination of many non-linguistic contents in speech signal.

In this chapter the correlation of certain spectral parameters and speaker's emotional states are investigated. It has been reported that the spectral parameters vary with different emotions. A detailed study of all the spectral parameters can provide information on their dependence on different emotions. Defining and valuation of the parameters that are relevant for recognizing emotional speech attributes is the newest element of research in speech scientific discipline and also qualitative improvement in speech technologies. It is clear that we are at the beginning of the process of research of the one complex scientific field that will open a lot of new segments in many scientific disciplines.

8. References

Ahmadi, F. & McLoughlin, I. (2010). The Use of Low-Frequency Ultrasonics in Speech Processing, In: *Signal Processing*, Sebastian Miron, pp. 503-528, InTech, Retrieved from: http://www.intechopen.com/articles/show/title/the-use-of-low-frequency-ultrasonics-in-speech-processing

Amir, N. (2001). Classifying Emotions in Speech: A Comparison of Methotds, *Proceedings of 7th European Conference on Speech Communication and Technology EUROSPEECH*, pp.127-130, Holon, Israel, September 2001

Amir, N. & Ron, S. (1998). Toward an Automatic Classification of Emotions in Speech, *Proceedings of the 4th Int. Conf. on Spoken Language Processing*, Sydney, Australia, Nov-Dec 1998

Bagchi, S. & Mitra, S.K. (1995). An efficient algorithm for DTMF decoding using the subband NDFT. *IEEE Int. Symp. On Circuits and Systems*, Vol. 3, (May 1995), pp. 1936-1939, ISSN 0271-4310

Bou-Ghazale, S.E., & Hansen, J.H.L. (2000). A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress. *IEEE Trans. on Speech and Audio Processing*, Vol. 8, No. 4, (July 2000), pp. 429-442, ISSN 1063-6676

Cowie, R.; Douglas-Cowie, E.; Tsapatsoulis, N.; Votsis, G.; Kollias, S.; Fellenz, W.; et al (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, Vol. 18, No. 1, (January 2001), pp. 32-80, ISSN 1053-5888

Delic, V.; Secujski, M.; Jakovljevic, N.; Janev, M.; Obradovic, R. & Pekar, D. (2010). Speech Technologies for Serbian and Kindred South Slavic Languages, In: *Advances in Speech Recognition*, Noam Shabtai, pp. 141-164, Sciyo, Retrieved from:

http://www.intechopen.com/articles/show/title/speech-technologies-for-serbian-and-kindred-south-slavic-languages

Dellaert, F.; Polzin, T. & Waibel, A. (1996). Recognizing Emotion in Speech, *Proceedings of Conf. on Spoken Language Processing ICSLP*, pp.1970-1973, Philadelphia, PA, October 1996

Dutt, A. (1991). A Fast Algorithm for the Evaluation of Trigonometric Series, *YALE UNIV NEW HAVEN CT DEPT OF COMPUTER SCIENCE*, January 1991

Felder,M. D.; Mason,J. C. & Evans, B.L. (1998). Efficient ITU-Compliant Dual-Tone Multiple-Frequency Detection Using the Non-Uniform Discrete Fourier Transform. *IEEE Signal Processing Letters*, Vol. 3, No. 7, (July 1988), pp. 160-163, ISSN 1070-9908

Flanagan, J. (May 1972). *Speech Analysis Synthesis and Perception*, Springer, Verlag, ISBN-13 978-0387055619, Berlin

Goertzel, G. (1958). An algorithm for the evaluation of finite trigonometric series. *American Mathematics Monthly*, Vol. 65, No. 1, (January 1958), pp. 34-35, ISSN 0002-9890

Hansen, J.H.L; Swail, C.; South, A.J.; Moore, R.K.; Steeneken, H.; Cupples, E.J.; Andreson, T.; Vloeberghs, C.R.A.; Trancoso, I. & Verlinde, P. (2000). The impact of speech under 'stress' on military speech technology. *NATO Res. Technol. Org. RTO-TR-10, AC/323 (IST) TP/5 IST/TG-01*, (March 2000), ISBN 92-837-1027-4

Ipsic, I. & Martincic-Ipsic, S. (2010). Croatian Speech Recognition, In: *Advances in Speech Recognition*, Noam Shabtai, pp. 123-140, Sciyo, Retrieved from: http://www.intechopen.com/articles/show/title/croatian-speech-recognition

Kienast, M. & Sendlmeier, W.F. (2000). Acoustical Analysis of Spectral and Temporal Changes in Emotional Speech, *Proceedings of the ISCA ITRW on Speech and Emotion*, pp. 92-97, Newcastle, Northern Ireland, UK, September 2000

Kiser, E. (2005). Digital Decoding Simplified Sequential Exact-Frequency Goertzel Algorithm. *CIRCUIT CELLAR*, No. 182, (September 2005), pp. 22-26, ISSN 1528-0608

Lee, C. & Narayanan, S. (2005). Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, Vol. 13, No. 2, (March 2005), pp. 293-303, ISSN: 1063-6676

Morrison, D.; Wang, R. & Liyanage C.D.S. (2007). Ensemble Methods for Spoken Emotion Recognition in Call-centres. *Speech Communication*, Vol. 49, No. 2, (February 2007), pp. 98-112, ISSN 0167-6393

Pekar, D.; Miskovic, D.; Knezevic, D.; Vujnovic-Sedlar, N.; Secujski, M. & Delic, V. (2010). Applications of Speech Technologies in Western Balkan Countries, In: *Advances in Speech Recognition*, Noam Shabtai, pp. 105-122, Sciyo, Retrieved from: http://www.intechopen.com/articles/show/title/applications-of-speech-technologies-in-western-balkan-countries

Petrushin, V.A. (2000). Emotion recognition in speech signal: Experimental study, development, and application, *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, October 2000

Ramamohan, S. & Dandapat, S. (2006). Sinousoidal Model-Based Analysis and Classification of Stressed Speech. *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 14, No. 3, (May 2006), pp. 737-746, ISSN 1558-7916

Scherer, K.R. (2003). Vocal Communication of Emotion: A Review of research paradigms. *Speech Commun.*, Vol. 40, No.1, (April 2003), pp. 227-256, ISSN 0167-6393

Ser, W.; Cen, L. & Yu. Z.L. (2008). A Hybrid PNN-GMM Classification Scheme for Speech Emotion Recognition, *Proceedings of the 19th International Conference on Pattern Recognition (ICPR)*, ISBN 978-1-4244-2175-6, Florida, USA, December 2008

Shafi, I.; Ahmad, J.; Shah, S.I. & Kashif, F. M. (2009). Techniques to Obtain Good Resolution and Concentrated Time-Frequency Distributions: A Review. *EURASIP Journal on Advances in Signal Processing*, Vol. 2009, Article ID 673539, 43 pages, 2009.doi:10.1155/2009/673539, ISSN 1687-6172

Sigmund, M. (2008). Automatic Speaker Recognition by Speech Signal, In: *Frontiers in Robotics, Automation and Control*, Alexander Zemliak, pp. 41-54, InTech, Retrieved from: http://www.intechopen.com/articles/show/title/automatic_speaker_recognition_by_speech_signal

Sigmund, M. (2009). Information Mining from Speech Signal, In: *Recent Advances in Signal Processing*, Ashraf A Zaher, pp. 297-319, InTech, Retrieved from: http://www.intechopen.com/articles/show/title/information-mining-from-speech-signal

Tao, J.; Kang, Y., & Li A., (2006). Prosody Conversion From Neutral Speech to Emotional Speech. *IEEE Trans. on Audio, Speech, and Language Processing*, Vol. 14, No. 4 (July 2006), pp. 1145-1154, ISSN 1558-7916

Tomas,B. (2006). Recognition of Linguistic and Nonlinguistic Information Contents in Speech Signal (in Croatian), *Proceedings of the 6th International Conference on Telecommunication BIHTEL*, Sarajevo, Bosnia and Herzegovina, October-November 2006

Tomas, B. & Obad M. (2008). Speech Signal Analysis by Goertzel Algorithm (In Serbian), *Proceedings of 16th Telecommunications Forum TELFOR 2008*, ISBN 978-86-7466-337-0, Belgrade, Serbia, November 2008

Tomas, B. & Obad M. (2009). Formant Vowel Structure Tracking by Goertzel Algorithm, *Proceedings of 4th International Conference on Digital Telecommunications*, ISBN 978-0-7695-3695-8, Colmar, France, July 2009

Tomas, B.; Maletić, M. & Obad, M. (2007a). Recognition and Implementation of Emotions in Speech Communications Model, *Proceedings of the 3rd Congress of the Alps Adria Acoustics Association*, JOANNEUM RESEARCH, Graz, Austria, September 2007

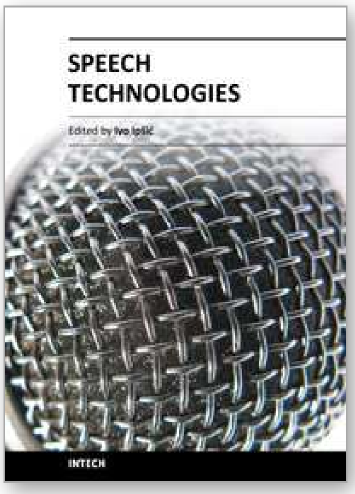
Tomas, B.; Maletić, M. & Raguž, Z. (2007b). Influence of Emotions to Pitch Harmonics Parameters of Vowel /a/, *Proceedings of the 49th International Symposium ELMAR-2007,* ISBN 978-953-7044-05-3, Zadar, Croatia, September 2007

Ververidis, D. & Kotropoulos, C. (2006). Emotional speech recognition: resources, features, and methods. *Speech Communication*, Vol. 48, No.9, (Sep. 2006) pp. 1163-1181, ISSN 0167-6393

Vojnović, M. (2004). Influence of Emotional State of the Speaker to Formant Structure of Vowel /a/ (in Serbian), *Proceedings of the 5th Congress DOGS*, Sombor, Serbia, September 2004

Wiliams, C.E. & Stevens, K.N. (1972). Emotions and Speech: Some Acoustical Correlates, *J.Acoust. Soc. Amer.*, Vol. 52, No. 4, (October 1972), pp. 1238-1250, ISSN 0001-4966

Zhou, G.; Hansen, J.H.L. & Kaiser, J.F. (2001). Nonlinear Feature Based Classification of Speech Under Stress. *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 3, (March 2001), pp. 201-216, ISSN 1063-6676



Speech Technologies
Edited by Prof. Ivo Ipsic

ISBN 978-953-307-996-7
Hard cover, 432 pages
Publisher InTech
Published online 23, June, 2011
Published in print edition June, 2011

This book addresses different aspects of the research field and a wide range of topics in speech signal processing, speech recognition and language processing. The chapters are divided in three different sections: Speech Signal Modeling, Speech Recognition and Applications. The chapters in the first section cover some essential topics in speech signal processing used for building speech recognition as well as for speech synthesis systems: speech feature enhancement, speech feature vector dimensionality reduction, segmentation of speech frames into phonetic segments. The chapters of the second part cover speech recognition methods and techniques used to read speech from various speech databases and broadcast news recognition for English and non-English languages. The third section of the book presents various speech technology applications used for body conducted speech recognition, hearing impairment, multimodal interfaces and facial expression recognition.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821