HAVE 2007 – IEEE International Workshop on
Haptic Audio Visual Environments and their Applications
Ottawa – Canada, 12-14 October 2007

Audio-Haptic Feedback in Speech Processing

Zygmunt Ciota
Department of Microelectronics and Computer Science, Technical University of Lodz,
Al. Politechniki 11, 90-924 Lodz, Poland
E-mail: ciota@dmcs.pl

Abstract *– The main goal of the paper is to achieve better human communication and interaction during conversation process by using a supervising of emotional states and improving voice quality. Therefore, the proposed approach should be also very helpful in the case of vocal tract illness for monitoring of treatment process. Since haptic feedback can nowadays operate by using different sense of touch, like kinesthetic, tactile, cutaneous or force feedback, then such feelings can be also helpful directly in medical treatments as the supplementary method.*

Keywords *– Audio-haptic interaction, Speech processing, Glottis excitation, Emotion recognition*

I. INTRODUCTION

The paper presents an attempt of speech feature's analysis and its possible application in haptic-audio interfaces oriented on medical applications. Mixed time- and frequency-domain analysis of voice signal gives a big number of feature vectors. Currently, such vectors are used generally in speech synthesis, speech recognition or speaker identification. It seems however, that some of features can be used in haptic interfaces (technologies) creating, together with audio information, a multimodal environment [4, 5, 8].

Therefore, the possibilities of modern speech processing should be first discussed. One of the most important tasks is a proper definition of feature parameters. Since usually impediments of speech depends strongly on emotional state of a speaker, a proper control of emotions plays a critical role in the treatment process. The speaking fundamental frequencies and time-energy distribution parameters permit to create emotion recognition system. The possibilities of different realization of decision algorithm for emotion classification have been also discussed. Researches concerning psychological and neurobiological analysis are also important and can improve the efficiency of intelligent human-machine interface. Unfortunately, it is impossible to find sharp boundaries between different emotions. Moreover, because a lot of different groups of scientists are involved in this topic, it is also impossible to obtain the final agreement concerning the number of emotions and a homogenous definition of them.

As a background, resulting from common behaviors of human being, we can take into account: joy, anger, sadness, disgust, fear, surprise and neutral state. Of course, for better understanding of mutual human relations it would be necessary to expand the list significantly, but for engineering purpose this background seems to be sufficient. Moreover, if our goal is focused on the correctness of emotion recognition, the number of emotions is inversely proportional to the recognition efficiency. The simplest analysis of only two emotions: negative and on-negative, can be sometimes very useful, especially in medical applications.

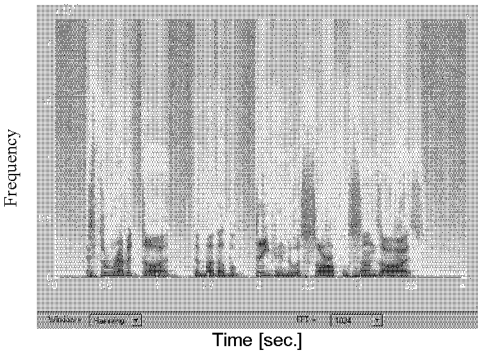

Fig. 1. An example of spectrogram

The calculation of the speech features is rather complicate and generally is based on time-frequency diagrams presented as spectrograms. An example of spectrogram, calculated using Hamming window and fast Fourier transform, is shown in Fig. 1.

II. EMOTION RECOGNITION

The proper choice of speech features has very significant influence for an efficiency of emotion recognition. The most important and useful features can be gathered into four main groups: the long-term spectrum (LTS), the speaking fundamental frequencies (SFF), the time-energy distribution (TED) and vowel formant tracking (VFT) [1, 2, 3]. Different emotions can be presented in multidimensional space as a function of the above features. An example of such function is shown in Fig. 2, where you have two-dimensional space describing by two axes: "Energy" and "Quality". If the number of space dimensions goes up, the situation becomes more and more complicate, and different emotion spaces can overlap. In such a case increasing of the features can even decrease recognition efficiency.

In our method we applied the features describing speaking fundamental frequency F_0 and time-energy distribution of the voice. Roughly speaking, TED is responsible for the energy and SFF for the quality of the speech. The efficiency of SFF depends on a processing system, which has to track and extract the fundamental frequencies precisely. As the results,

978-1-4244-1571-7/07/$25.00 ©2007 IEEE

statistical behaviors of these frequencies have to be also included and fortunately, these frequencies are very sensitive to different emotions of the same speaker, like anger, joy, sadness or boredom.

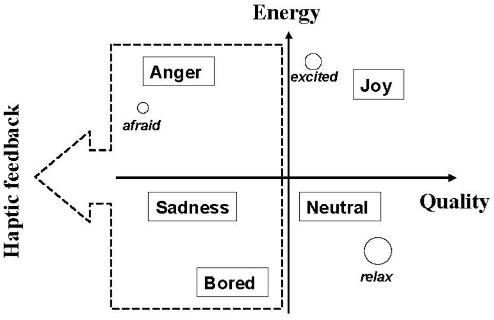

Fig. 2. Emotions in two-dimensional space described by energy and quality of the speech

The input signal corresponding to glottal waves can be simulated using different models. An example of useful model based on electro-mechanical equivalents of human glottis is shown in Fig. 3. Such model can be used as an input of whole vocal tract including models of nasal and mouth tracts, according to Fig. 4.

In the case of fundamental frequency calculation, two basic methods are available: autocorrelation and cepstrum method. The first permits to obtain precise results, but we discovered that additional incorrect glottis frequencies have been created. Additionally, program indicates some glottis excitations during breaks between phones and in silence regions. Therefore, in this method it is necessary to apply special filters to eliminate all incorrect frequencies.

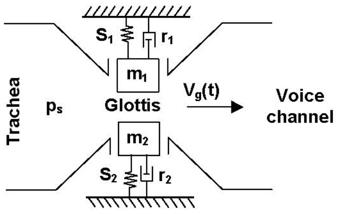

Fig. 3. Modeling of glottis behavior

Another method bases on cepstrum analysis. In this transform the convolution of glottis excitation and vocal tract is converted, first to the product after Fourier transform, separated them finally as the sum. In our method we use cepstrum analysis as less complex, especially when we applied modulo of cepstrum by using modulo of Fourier transform. The following values of glottis frequency F_0 have been taken into account: minimum and maximum values: $F_{0\text{-}minimum}$ and $F_{0\text{-}maximum}$, the mean value $F_{0\text{-}mean}$, and the range $F_{0\text{-}range}$. The parameters describing time-energy distribution have been calculated using fast Fourier transform and dividing speech utterances into 20 ms slots. As the result we obtained eight values of the energy for the following frequency ranges: 0-400Hz, 400-800Hz, 800-1500Hz, 1500-2000Hz, 2000-3500Hz, 3500-5000Hz and 5000-8000Hz.

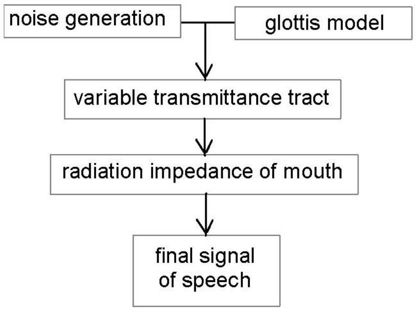

Fig. 4. Modeling of the whole vocal tract

The process of emotion recognition consists of two main parts: emotion teaching and appropriate recognition. During the teaching process one can create the base of parameters for all emotions. The comparison of current voice with the base gives the answer concerning the emotional state of examined utterance. The comparison process and the final decision are based on two classifiers: nearest mean (NM) and nearest neighbour (NN). The decision process can be optimized using different distances and parameter weights. This part of method is very important and still open. Especially, in the present of low quality teaching materials, it would be necessary to applied probabilistic method and multilayer neural perceptrons [6].

III. FUNDAMENTAL FREQUENCY FEATURES

In the case of medical application haptic feedback should be helpful in diagnosis and treatment processes. Malfunction of pronunciation depends strongly on such emotional states like anger, fear, sadness and boredom, because of glottis signal deterioration. In the other words, the fundamental frequency F_0 becomes unstable. Information of such changes can be helpful for doctors for progress observation of glottis treatment. Moreover, such acoustic feedback should be helpful for patient himself, increasing his attempt to improve the pronunciation of the speech. The threshold, in which the frequency properties of the glottis signal make worse, can be obtained during learning process using database of patient's utterances. From technical point of view the haptic feedback can be realized using a ring attached to a hand and containing a vibrator. The frequency of vibration should be dependant on the intensity of negative emotions.

Table 1. F_0 parameters for male voice

	F_0-*mean*	F_0-*max*	F_0-*min*	F_0-*range*
Anger	172 Hz	228 Hz	120 Hz	108 Hz
Neutral	106 Hz	142 Hz	94 Hz	48 Hz

Table 2. F_0 parameters for female voice

	F_0-*mean*	F_0-*max*	F_0-*min*	F_0-*range*
Anger	201 Hz	232 Hz	106 Hz	126 Hz
Joy	209 Hz	240 Hz	165 Hz	76 Hz

The system require several time-to-frequency transforms, but it is necessary assure real-time action. Therefore, the implementation of the proposed algorithms using hardware-software system, including mixed analog-digital approach, should improve the speed and the quality of proper action. Application of mixed digital-analog realization to the design process of sound processors may be better in comparison with purely digital solution and very often we can achieve better results, decreasing the chip surface and increasing the speed parameter of the system.

Extraction of fundamental frequency parameters for male and female voices is shown in Table 1 and Table 2, respectively. It is very easy to observe high sensitivity of such parameters for emotional state of the speaker. The range of F_0 is equal to 126 Hz for anger speech of a women, while in the case of joy speech the corresponding range decreases to 76 Hz. Similarly, high sensitivity of fundamental frequency parameters can be observed for angry and neutral speech of a man (see Table 1).

IV. MIXED HARDWARE-SOFTWARE APPROACH

The design process of integrated circuits is focused on the following two main directions: to scale down the dimensions of the transistors, and to incorporate as many building blocks as possible in a single chip. It means that modern CMOS technology requires low-voltage power supply.

VLSI techniques usually demand high-performance A/D and D/A converters, because digital circuits have to be interconnected to the real and in most cases analog world. As a consequence, almost every fully integrated system contains some specific mixed blocks. In many applications analog to digital converter remains the most important component of analog circuits. The design process of high performance converters is complicated and time consuming. The performance depends also no the process technology, therefore it is necessary to obtain a compromise between the precision, power consuming and the cost of the chip.

The design process based on a current-mode approach can be applied in standard CMOS technology. Analog integrated circuits can be divided in time continuous and sampled-date techniques. In the sampled-date group switched-current circuits are the most important, because this technique requires only a standard digital CMOS process. Such circuits use MOS transistors as storage elements to provide analog memory capability. Using current mirror principles, a voltage is sampled onto the gate of MOS transistor and held on its gate capacitance. An example of the family of frequency characteristics for switched current filter, made as integrated CMOS prototype, is presented in Fig. 5.

Current mode realizations of analog components permit to decrease chip area significantly. As the most important analog realizations we can mention filters and analog-to-digital converters. Comparison of the different design methods permits to choose the best one with regard to the kind of technology in which the analogue-digital system should be integrated. Preliminary computer simulations are also very important and can be helpful to obtain low sensitivity to variations in technological parameters. Design of audio processing systems, including computer speech circuits is inherently a complex task involving human expertise as well as aids intended to accelerate the process.

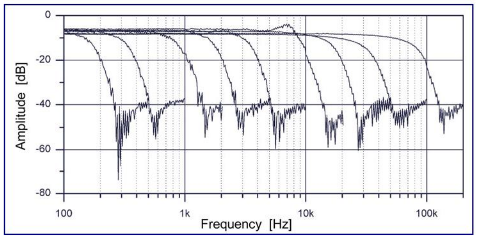

Fig. 5. Amplitude characteristics of switched current filter for different clock frequencies

Furthermore, such efficient system has to have real-time capabilities, so the hardware-software co-design permits to achieve low cost and high-speed performances. While microcontrollers and microprocessors are inherently digital components, some functions can be executed in analog or digital form. An example of simple speech recognition system based on ARM processor is shown in Fig. 6.

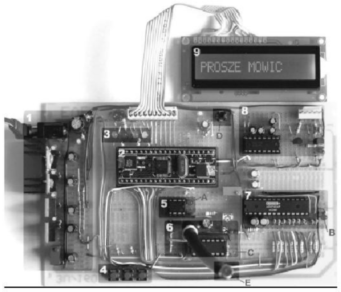

Fig. 6. An example of speech processing prototype using ARM processor

The general trend is towards digital solutions, which guarantee high density and easy design. Available digital design tools, for example Cadence environment, can automatically convert the logic scheme to the chip layout. We expect that hardware realization of some components can be useful for smart real-time applicable recognition system.

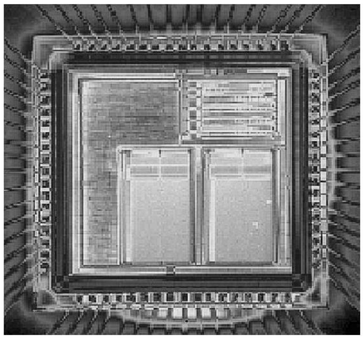

Fig. 7. Microphotograph of audio processing chip (CMOS technology)

The mentioned mixed signal system for audio signal processing was built for prototype purposes and preliminary researches over integration of different parts on a monolithic silicon die (see Fig. 7). The bulk CMOS technology with all advantages (economic) an disadvantage (quality) was used. The future belongs to SOI CMOS technology, especially in area of System-on-Chip integration. Better substrate noise isolation, lower threshold voltages of MOS transistors and lower power dissipation are key aspects in this case.

Taking into account the above remarks, a system of audio processing has been designed and performed in CMOS technology [9]. The prototypes contain high precision CMOS integrated filters with low sensitivity to technological mismatches. The proposed laboratory stand permits to obtain accurate current measurements for such circuits. The prototype contains also digital memory to control all functions of audio signal processing, including software radio capabilities [9].

The digital part has driving purposes, consisting from analogue part configuration system, and also software radio signal processing. The novel method of FM signal demodulation was introduced based on asynchronous digital circuit, using standard CMOS digital cells. Direct pulse counting was used for FM detection, the obtained samples was processed by the fixed-point arithmetic circuit. Digital sigma-delta modulator was used as additional output interface. The analogue part has been designed as switched-current reprogrammable cells. The coefficients for this part are loaded from the digital part.

The presented results show new possibilities of the system integration level. Combining analogue SI circuits and digital signal processing at the same chip we reduce overall costs giving more robust and flexible system. The System-on-Chip allows to build complete structure for up to date multimedia purposes with low power dissipation and modest costs in high volume production series that is very important for many of the modern portable devices.

Presented in this paper reprogrammable SI cells could be treated as first step in exploration of the very promising technique, often proclaimed as the antecessor of widely used SC discrete-time processing. Lack of the industrial proposals is a gap that potentially could be filled by reprogrammable SI circuits. Higher usage of SI could reveal and improve its negative features, which might develop better solution for discrete-time signal processing.

Current works are focused on the integration possibilities of haptic interface, together with speech processing system, in a single chip.

V. CONCLUSIONS

Detailed analysis of the recognition efficiency leads us to the following conclusions. Using mixed analogue-digital methods we can obtain better acoustic signal processing results, like compression, identification and recognition. Exploring more general problem, the improvement of the method has been depended on the voice quality and environment conditions.

The proper and fast system of speech parameters extraction is a difficult task, because the quality of speech processing depends strongly on different environments. It is also necessary to take into account, that different words have different recognition difficulties, therefore the precision of extraction depends on a spoken text. We expect that hardware realization of some components, e.g. predictive filters, neural network units, can be useful for smart real-time applicable predictors.

The implementation of the proposed algorithms as hardware-software system, including mixed analog-digital approach, should improve the speed and also the quality and resolution of the system. The presented prototypes contain high precision CMOS integrated filters with low sensitivity to technological mismatches. The proposed laboratory stand permits to obtain accurate current measurements for such circuits.

We expect that development of our existing speech processing systems by adding haptic interfaces, can enlarge the area of possible application in medical treatment and diagnosis of speech abnormalities.

ACKNOWLEDGMENT

The research effort is sponsored by the grant of Polish Ministry of Education and Science No. 3T11B 027 29

REFERENCES

[1] Z. Ciota: “Speaker Verification for Multimedia Application”, IEEE International Conference on Systems, Man and Cybernetics, October 10-13, 2004, Hague, Netherlands, pp. 2752-2756

[2] Progress in speech synthesis, edited by J. Santon et al., Springer, New York 1996

[3] Z. Ciota: “Emotion Recognition on the Basis of Human Speech”, ICECom-2005, 18th International Conference on Applied Electromagnetics and Communications, 12-14 October 2005, Dubrovnik, Croatia, pp. 467-470

[4] Charlotte Magnusson, Kirsten Rassmus-Gröhn: "Audio haptic tools for navigation in non visual environments", Proc. ENACTIVE05 2nd International Conference on Enactive Interfaces, Genoa, Italy, November 17-18, 2005

[5] G. Nikolakis, I. Tsampoulatidis, D. Tzovaras and Michael G. Strintzis: Haptic Browser: "A Haptic Environment to Access HTML Pages", SPECOM’2004: 9th Conference Speech and Computer, St. Petersburg, Russia, September 20-22, 2004

[6] Chulhee Lee, Donghoon Hyun, Euisun Choi, Jinwook Go, Chungyong Lee: “Optimizing Feature Extraction for Speech Recognition”. IEEE Trans. Speech and Audio Processing, no 1, January 2003, pp. 80-87

[7] Z. Ciota: “Improvement of speech processing using fuzzy logic approach”, Proc. Joint 9th IFSA Word Congress and 20th NAFIPS International Conference, July 25-28, 2001, Vancouver, Canada, pp. 727-731

[8] B.D. Adelstein, D.R. Begault, M.R. Anderson1, E.M. Wenzel: "Sensitivity to Haptic-Audio Asynchrony", Proceedings, 5th International Conference on Multimodal Interfaces, ACM, Vancouver, Canada, 2003, pp. 73-76

[9] Rominski A., Ciota Z., Napieralski A: "Audio Signal Processing Using Mixed Hardware-Software Approach", Proc. Nanotechnology Conference and Trade Show, NSTI Nanotech, Boston, Massachusetts, USA 2006, pp. 63-65