Adam DĄBROWSKI, Paweł PAWŁOWSKI, Radosław WEYCHAN, Andrzej MEYER, Marek PORTALSKI, Agata CHMIELEWSKA, Tomasz JANIAK

Poznań University of Technology, Chair of Control and System Engineering, Division of Signal Processing and Electronic Systems

Real-time watermarking of one side of telephone conversation for speaker segmentation

***Abstract**. The paper presents a digital signal processor (DSP) based system for segmentation of speakers of a telephone conversation. The TMS320C6713 DSP by Texas Instruments in real-time watermarks one interlocutor voice and therefore precise segmentation of both conversation sides is made on a PC without any speaker recognition techniques. The authors also solved the problem of data blocks synchronization and beats caused by differences in the digital-to-analog and the analog-to-digital sampling clock frequencies.*

***Streszczenie.** Artykuł prezentuje, zrealizowany na procesorze sygnałowym, system do segmentacji mówców rozmowy telefonicznej. Użyto procesora TMS320C6713 firmy Texas Instruments, który podczas rozmowy oznacza znakiem wodnym jednego z rozmówców. Umożliwia to późniejszą separację mówców bez użycia algorytmów ich rozpoznawania. Autorzy dodatkowo rozwiązali problemy związane z synchronizacją bloków danych i dudnieniami wywołanymi różnicą częstotliwości zegarów taktujących przetworniki analogowo-cyfrowe i cyfrowo-analogowe. (**Wprowadzanie w czasie rzeczywistym znaku wodnego do sygnału jednej strony rozmowy telefonicznej w celu segmentacji mówców**)*

Keywords: DSP, watermark, speaker segmentation, DWT
Słowa kluczowe: procesor sygnałowy, znak wodny, segmentacja mówców, dyskretna transformata zafalowaniowa

Introduction

An idea of the presented system arose from the need for a reliable speaker segmentation during a typical phone call, (e.g., to the emergency services numbers). The inserted watermark to the emergency telephone operator voice makes subsequent conversation sides separation possible with no additional necessity of the speaker recognition.

The speakers recognition algorithms, although deeply investigated, cannot guarantee 100% reliability of the speaker segmentation. Additionally they are often computationally complex, thus they need strong computing devices or long times (typically an off-line processing).

The authors proposed a mixed hardware/software solution with the use of a watermarking technology. The system is assumed to be used in a case of accessible telephone set of one of the conversation sides. A DSP (*digital signal processor*) adds in real-time a not hearable or an almost not hearable watermark to the voice of the accessible speaker. During inserting the watermark the signal distortion is low the voice delay is not disturbing. During a conversation, the latter parameter, namely the delay, cannot exceed a value, for which the time shift of the watermarked voice is annoying or even obstructing the conversation.

On one hand the watermark should not be hearable, but, on the other hand, it should not be sensitive to the analog telephone line limitations (e.g. reduced passband, low dynamic range, noise, non-linear distortions, etc.). The watermark must also be prepared and added in a way, that all important information will be recorded by a typical telephone line recorder [1].

The segmentation part, which relies on the detection of the watermark, operates off-line and processes files registered by the telephone line recorder. The recorder registers the call with the usage of a signal from the analog telephone line, so there are no additional information, that could support the watermark detection phase. In fact, the analog stage of the system brings a lot of problems: differentiation in voice levels and distortions, the reduced passband, no certainty of the exact speed of the analog-to-digital (A/D) and the digital-to-analog (D/A) conversions, no data about location of the blocks (the digital part of the system processes signals in blocks, thus they should be synchronized). All problems that affected each step of the processing were deeply analyzed and successfully solved by the authors.

The paper presents an overall solution, which is prepared to support the emergency telephone offices, as such telephone lines are often misused and abused. For this reason, it is sometimes required to monitor and identify the callers. The main purpose of the system supporting the work of law enforcement is to extract useful information from the recordings of the emergency calls. The task of identification people must be supported by a system of segmentation of the speakers, in order to define and extract the borders of the speech of both conversation sides.

In [2] the authors proposed a simple and effective method of inserting the watermark, which is a rectangular function multiplied with the signal. Extracted sentences uttered by a citizen, have led to speaker segmentation and removal from the recording the voice of an officer, in order to better identify the citizen. In the present paper we propose to insert the watermark in a set of the DWT (*discrete wavelet transform*) coefficients.

An important feature of our approach is that the segmentation task is performed without any knowledge of the speakers voice. This is a big difference between our paper and other publications. For example [3] deals with the problem of the speaker change detection by three parallel fused classifiers: multi layer perceptron, SVM (*support vector machine*) and a classifier based on the Mono-Gaussian statistical measure. Articles [4, 5] present speaker segmentation based on BIC (*Bayesian information criterion*) metric. Paper [6] presents an on-line speaker segmentation, but in this algorithm the acoustic models are also used. Speaker segmentation in the telephone conversation is proposed in [7]. A preliminary segmentation to hypothesize speaker turning points was used, then the clustering of the segment and re-segmentation to determine the speaker identity was applied. An analysis of the power in different frequency bands obtained by the DWT in continuous speech was shown in [8].

Real-time watermark generator – hardware

Figure 1 shows the watermarking system and Figure 2 presents the device without casing. The system inserts the watermark into one interlocutor voice (to the emergency phone operator's voice) during a telephone conversation. The detailed description of the system is given below. The watermarking process is realized in real-time by the TMS320C6713 digital signal processor (DSP) by Texas Instruments. A connection of the DSP to the analog telephone line is not trivial. It is shown as a block diagram in Fig. 3.

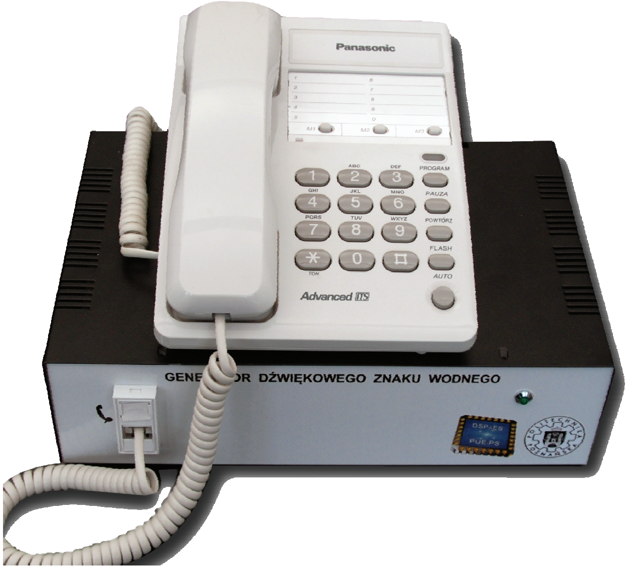

Fig.1. Real-time watermark generator as a telephone adapter

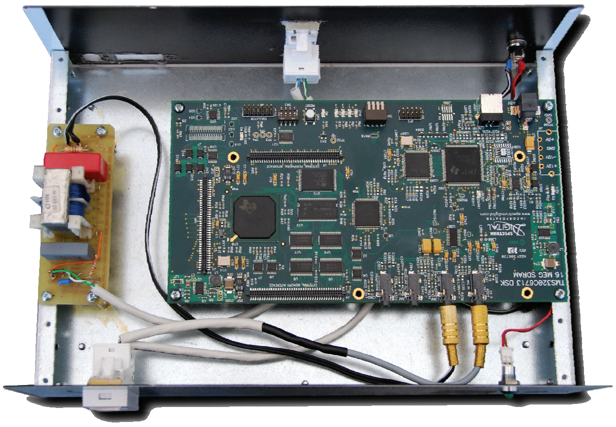

Fig.2. Real-time watermark generator (DSP inside)

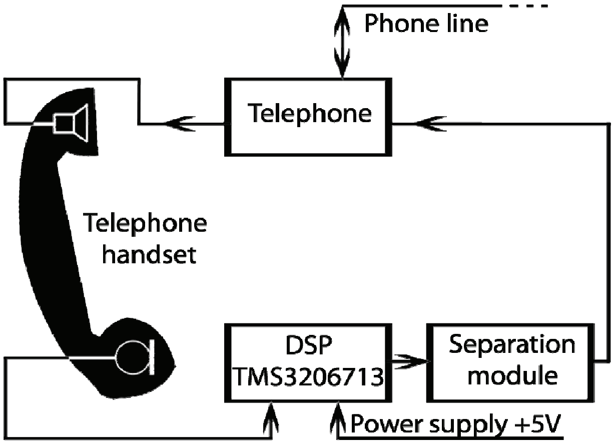

Fig.3. A block diagram of the watermarking system

The device is prepared as a telephone adapter and can work with any analog telephone with a typical wired receiver (handset). The handset is attached to the adapter by a commonly used RJ-9 modular connector, another RJ-9 socket (located on the rear side of the adapter) connects the DSP with the telephone. Because the microphone, that is located in the handset is of the electret type and is also galvanically separated from the telephone, it can be directly wired to the microphone input at the DSP board. The electret microphone must be powered and it is done by the analog interface on the main printed circuit board (PCB).

The DSP adds the watermark to the signal from the handset through an extra prepared separation module, it sends the modified operator voice to the telephone. The DSP card is externally powered with +5V DC.

The separation module (see the left side of Fig. 2) must ensure that the phone is galvanically separated from the DSP card. It is required due to the following reasons: the DSP is equipped only with an asymmetric analog output and the impedance of the output differs from the input of the telephone. While the analog telephone line is symmetric with respect to the ground, additional load to the ground (with resistance of several kΩ) causes strong and hearable interference to the telephone exchange. The scheme of the proposed separation module is shown in Fig. 4.

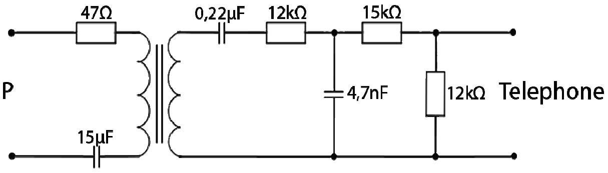

Fig.4. Schematic diagram of the separation module

An extremely important element of the module is the transformer (see Fig. 2 and 4). This element was extra wound: the primary winding has 210 turns, secondary 190 turns. The measured parameters of the transformer are given in Table 1. To check the properties of the transformer, the THD (*total harmonic distortion*) introduced by the transformer was measured. The results are shown in Fig. 5.

Table 1. Transformer parameters (measured)

Parameter	Unit	Value
Transmission	[-]	0.89
Main inductance	[mH]	61
Resistance of primary winding	[Ω]	3.2
Resistance of secondary winding	[Ω]	2.9
Cross-section of core (EE-shaped)	[cm^2]	1.1

The obtained results show that the distortion of the output signal is, as for the analog telecommunications equipment, small enough, to avoid hearable artifacts.

The primary winding of the transformer is connected to the DSP card with the series RC circuit. The capacitance is 15 µF. This capacitance together with the inductance of the primary winding form an LC circuit with the resonant frequency of 166 Hz, which is below the acoustic band of the phone (typically 300 Hz - 3,4 kHz). An additional resistor of 47 Ω lowers the Q-factor of the LC circuit and attenuates the currents in the transient states, protecting the output of the DSP card against overload.

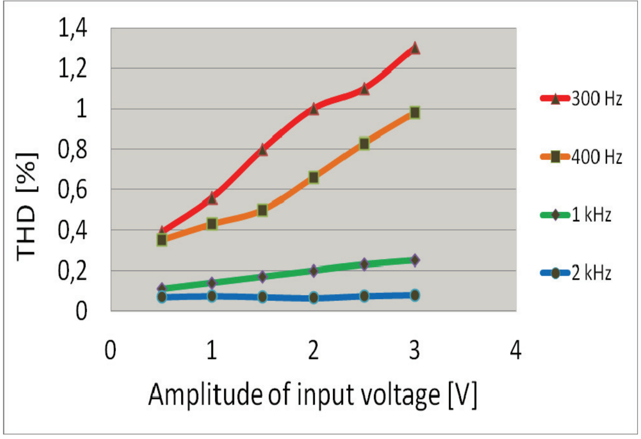

Fig. 5. Total harmonic distortion (THD) caused by the transformer

An RC circuit shown in Fig. 4 adjusts the transformer output to the microphone input of the phone. A parallel resistor of 12 kΩ forces the DC current through the microphone, necessary for its proper operation. Other resistors and a capacitor (with a capacity of 4.7 nF) form a low-pass filter (with the frequency limit equal to 4.6 kHz) that improves the action of the filter after the D/A conversion and cuts audible distortions. In addition, the system ensures the proper RC fit to the signal level between the transformer and the phone; a 0.22 μF capacitor separates the DC component.

Another element of the equipment is the DSP TMS320C6713 DSK type module, which realizes the watermark inserting algorithm. Technical specifications of this DSP module are as follows [9]:

- Texas Instrument's TMS320C6713 DSP operating at 225 MHz
- TLV320AIC codec with four 3.5 mm audio jacks (microphone, line-in, speaker, and line out)
- 2M x 32 on board SDRAM
- 512K bytes of on board flash ROM
- 3 expansion connectors (memory interface, peripheral interface, and host port interface)
- 4 user definable LEDs and 4 user definable DIP switches.

Real-time watermark generator – software

The watermark is placed in the operator's voice signal and it is inserted only while the operator speaks. A digital watermark in the audio signal should not provide any hearable distortion for listeners, but it needs to be reliably detectable in the marked parts [2]. On another side, as the watermark should be inserted in real-time, the procedure cannot be very complicated and, because of the delay limits, must enable block processing. Taking the above requirements into account, the authors decided to modify the signal in the DWT space [11]. In order to achieve the most effective masking of the watermark by the speech signal and a sufficient time resolution, the division into 256 uniform bands with the *Symlet 12* wavelet function has been used. Uniform bands constitute the so called wavelet packet transform. The *Symlet* is the most advanced version of wavelets and was selected from three types of the considered orthogonal wavelets: *Coiflets, Daubechies,* and *Symlet*. An order of the analysis filter, namely 12, is a compromise between efficiency of the analysis and computational complexity. Figure 6 shows sequences of the low- and the highpass filter coefficients associated with the used wavelet function [13].

An inserting of the watermark consists in putting a non zero constant instead of the 128-th coefficient in the DWT analysis block. In the opposite case (the speech signal without the watermark) this coefficient is zeroed. The 128-th coefficient has been chosen because of the best efficiency of speech masking in about 1.3 kHz band.

Figures 7 and 8 show block diagrams of the watermark inserting algorithm. The algorithm has been developed in the Matlab / Simulink environment. The operator's speech is sent from the microphone to the microphone input on the DSP board (cf. Fig. 3 and 7). After the A/D conversion, the one-dimensional signal sampled with 8000 Sps has a 16 bit resolution. During the conversation, the input signal is gained twice (6dB), which compensates for voltage drop in a passive element such as the D/A receiver coupling module (block *Gain* in Fig. 7). Then the speech signal is analyzed by the DWT algorithm with the *Symlet 12* wavelet function. A decision about the watermark insertion is taken in the *Gating* subsystem shown in Fig. 8. In every signal frame of 32 ms duration, the RMS (*root mean square*) of the current sequence of the DWT coefficient is calculated. The value is compared with the *Threshold*. The results of this comparison are stored in the shift register with the length of 15 (*Buffer* in Fig. 8). Next, the values from the register are sent to the OR gate in order to maintain the value "1", when short breaks in the signal occur (brakes between words or breathing). The amplitude of the watermark signal is set by a multiplier with an initial v*alue* equal to 0.2. The last step of the algorithm is the synthesis of all coefficient sequences. It is realized with the IDWT (*inverse DWT*). The results of the reconstruction are finally converted to an analog signal and sent by the separation module to the telephone.

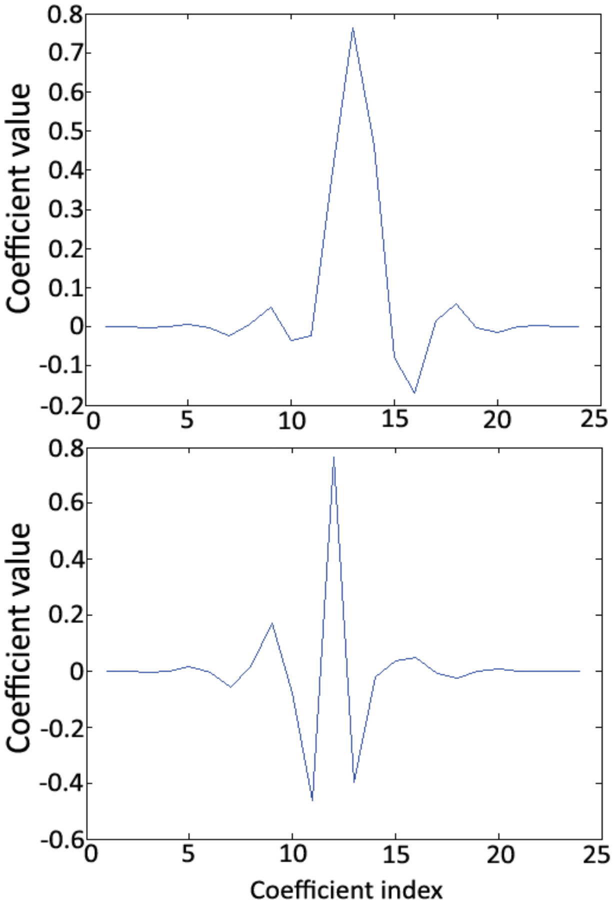

Fig. 6. Low- (top) and highpass (bottom) filter coefficients for "Symlet 12" packet

A hardware / software implementation of the described algorithm uses the DSP module and the link for CCS (*Code Composer Studio*) library for the Matlab / Simulink environment. This library allows to convert the designed schemes and the configuration blocks of the devices to C++ language, acceptable by the CCS compiler. To configure the external modules, it is necessary to set the sampling rate of the audio signal, the length of the input buffer and properties of the microphone and line inputs. These operations are performed by the ADC and DAC blocks (see Fig. 7). They operate synchronously, and by this means determine a precisely the same sampling rate for both of them. The sampling has been set to 8000 of 16-bit Sps (*samples per second*). After the A/D conversion the data are transmitted with the MCBSP (*multichannel buffered serial port*).

The memory map configuration is realized by the C6713DSK block (visible as a PC card in Fig. 7). The source code, variables, the input and output buffers are placed in the external SDRAM with the first address equal to 0x80000000 and the length of 0x00030000, while the system stack is placed in the (internal) IRAM from 0x00000000 to 0x00030000. The memory map of the C6713DSK module is summarized in Table 2 [9].

The algorithm of the DSP based real-time watermarking of one side of the telephone conversation has been tested in typical calls. It has been noticed, that the watermark is almost not hearable and it does not influence the call quality for both the recipient and the operator. The most critical parameters that control the data flow, including the length of the shift register maintaining continuity of the watermark inserting, have been optimized experimentally.

Speaker segmentation

A telephone call is recorded by the NetCRR – a typical telephone line recorder [1]. It registers non compressed wav-type files. The recorder stores 16-bit samples with the sampling rate equal to 8000 Sps. The speaker segmentation software operates on a PC and processes off-line the input files previously registered by the line recorder. An illustrative, previously watermarked, call waveform is depicted in Fig. 9(a). The parts with white background depict one interlocutor and the parts with grey background depict the second one. It is visible that due to different sensitivities of the telephones (in the experiments two meaningfully different phones were used) and different volumes of the speakers (this parameter is quite unpredictable) the levels of the call waveforms are somewhat dissimilar for both speakers. The watermarked speaker has a lower volume.

Table 2. C6713DSK Memory map

Address	C67x Family Memory Type	6713 DSK
0x00000000	Internal Memory	Internal Memory
0x00030000	Reserved Space or Peripheral Registers	Reserved or Peripheral
0x80000000	EMIF CE0	SDRAM
0x90000000	EMIF CE1	Flash
		CPLD (0x90080000)
0xA0000000	EMIF CE2	Daughter Card
0xB0000000	EMIF CE3	

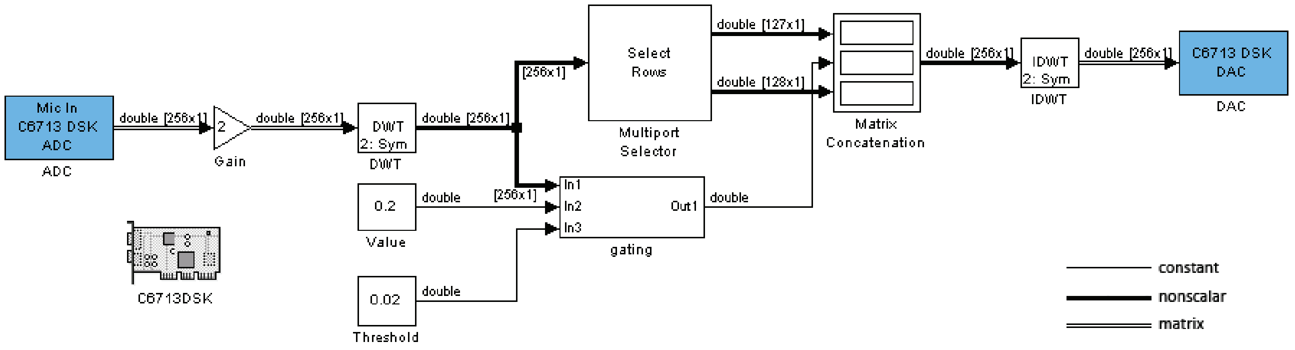

Fig. 7. Model-design block diagram of watermark inserting in phone call operator’s speech

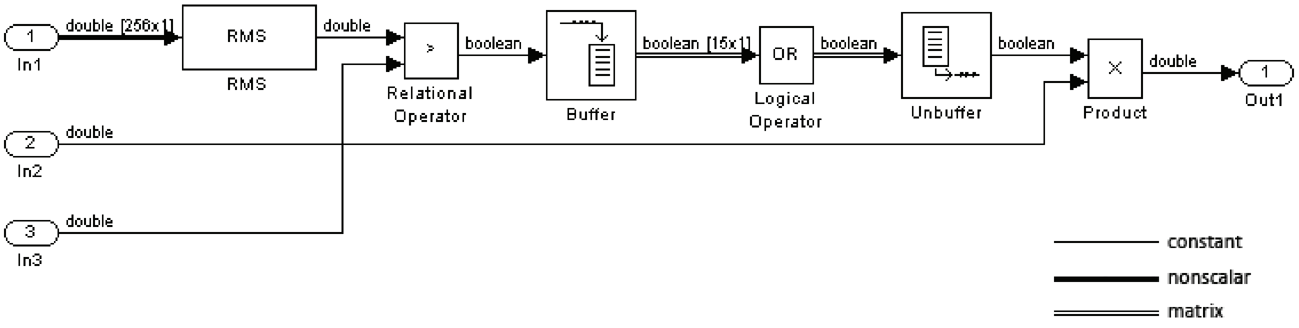

Fig. 8. Block diagram of speech signal gating procedure (subsystem “*gating*” in Fig.7)

The segmentation software in the beginning, analyzes the signal with the use of the DWT and selects the 128-th coefficient for each block of 256-samples (see Fig. 9(b)). A watermark presence means that the 128-th coefficient in the DWT analysis of a block is filled. If it is zeroed, it means no presence of the watermark. It is true for one speaker, but in fact, the second speaker can also fill this block by his voice (the analog telephone line has only one full-duplex channel for both speakers). The mentioned block represents a band located near to 1350 Hz. The band can be affected e.g. by non-linear distortions of the speaker voice or another unwanted source, that can appear in the telephone line.

The peaks visible in Fig. 9(b), after selection of the 128-th DWT coefficient, illustrate the problem. The watermark cannot be properly detected by a simple thresholding technique.

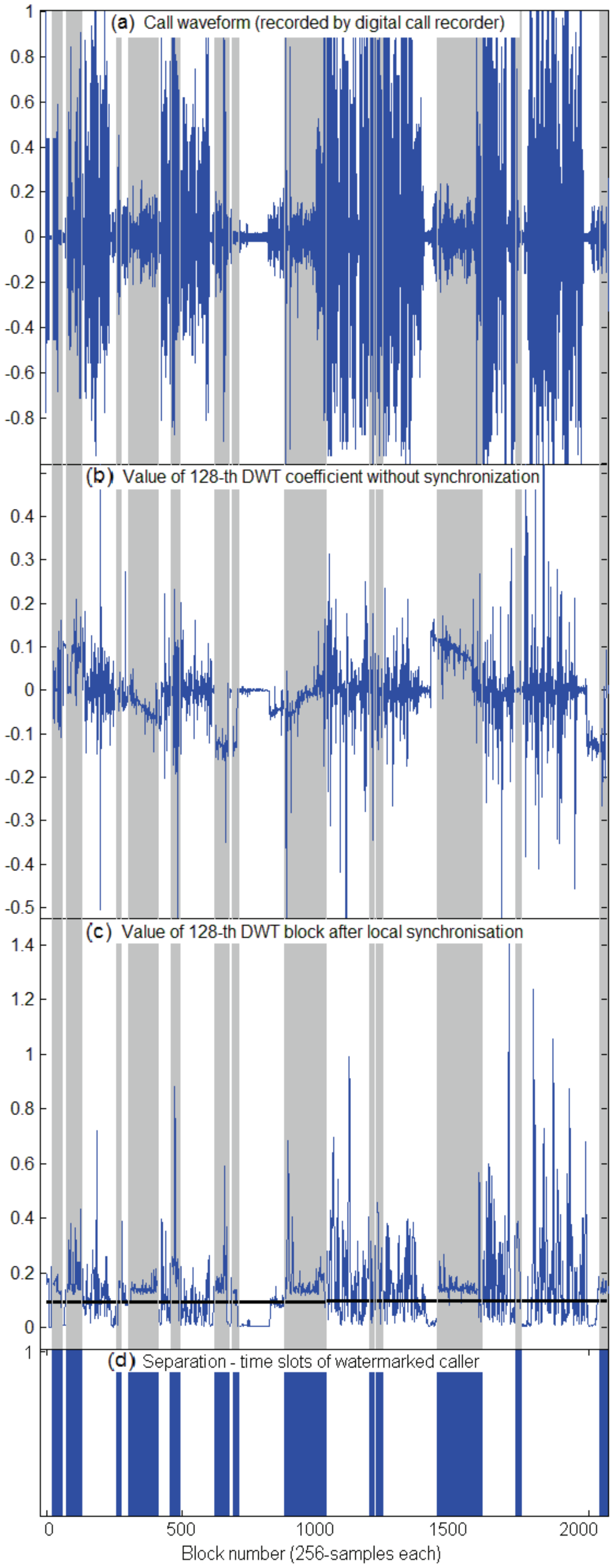

Fig. 9. Speaker segmentation: from top:
(a) Example call waveform (recorded by digital call recorder)
(b) Value of 128-th DWT coefficient without synchronization
(c) Value of 128-th DWT coefficient with synchronization (black line depicts the watermark detection level)
(d) Time slots of watermarked caller

Another harmful phenomenon is a series of rumbles caused by differences of the sampling clock frequencies controlling the D/A and A/D converters. It is visible after the watermark detection (cf., Fig. 9(b)). In the presented system both D/A (the converter is placed on the DSP board) and A/D (the converter is applied in the digital call recorder) use clocks with the frequency equal to 8000 Hz stabilized by crystal oscillators. Typical low-end crystal oscillators have accuracies of ±30 ppm. In case of our experiments they produced rumbles with a period of about 50 seconds (see the low frequency sine modulation in Fig. 9(b)). It is possible to calibrate the system by sending the watermark signal only (without speech signal) and calculate the FFT of it, but the differences between frequencies cause a reasonable size of the required FFT for this method of about a million. This is quite impractical, as it means about two minutes of the signal. In result, there are no possibilities to calibrate the system call by call. In addition it is not certain that the A/D and D/A conversion speeds will remain exact for a long time.

The next problem is synchronization of blocks. Both digital parts of the system (i.e. the DSP, that inserts the watermark and the PC software that detects this watermark) process signals in blocks. To obtain proper results they should be block-synchronized. The first idea is to perform the block synchronization once at the beginning of the call by maximization of the product at the detection stage. Unfortunately, due to differences of the frequency of sampling clocks, the synchronization will be actual for several seconds only.

All above problems have been solved by means of a local synchronization of blocks. The improved watermark detection procedure, for each processed block, looks for the best match of the block under the detection window by shifting it in both the left and the right hand side, up to a half of the previous and the next block. It gives the brilliant result depicted in Fig. 9(c). A threshold at a given level (here equal to 0.1) searches the time slots above the threshold.

Additionally, the detection procedure reduces the glitches with the use of the fact that the shortest watermark has been inserted for at least 15 consecutive blocks (i.e. about 0.5 s). Also, if between the speakers' voice there is a silence, it is cut by additional RMS (*root mean square*) thresholding.

Finally, the prepared software produces time markers (blue boxes in Fig. 9(d), separates the operator and the interlocutor voices, cuts the original single-channel conversation signal into two signals, and put them into separate channels, with one speaker (or more precisely conversation side) each. The signal prepared in this way is ready for further processing, e.g., for transcription.

Conclusions

In this paper a system for the speaker segmentation of the registered phone conversation has been presented, based on the watermark insertion to the signal of one side of the conversation.

Thus the watermark is inserted into the signal during the speech of one of the speakers – in the considered case to the operator's voice, receiving a phone call to the emergency services number. The watermarking task is performed by the properly interfaced DSP, so the quality of processing is high. Here, the quality is defined in the sense of suppression of audible distortions like: the watermark itself, noise, delay, THD, etc.

In the watermark detection stage, an important feature is that the speaker segmentation task is performed without any knowledge of the speakers' voice. Additionally, an acoustic background, mixed with the speech signal does not affect the quality of segmentation.

As the system consists of typical devices for the considered application, e.g. telephones, analog telephone line, exchange, and digital voice recorder, it brings a lot of inaccuracies to the signal processing. A simple, but effective procedure for the local synchronization of the speech signal blocks leads to the improved, amazingly stable, and reliable watermark detection procedure.

This work was supported by INDECT and DS 2012 projects.

REFERENCES

[1] DGT, NetCRR digital call recorder. Catalogue card. Available at: www.dgt.com.pl

[2] Dąbrowski A., Weychan R, Meyer A., Chmielewska A., Segmentacja mówców w rozmowach telefonicznych na podstawie znaku wodnego, *Elektronika (LII),* nr 5/2011, (2011), 98-102

[3] Ouamour S., Sayoud H., Guerti M., Speaker Segmentation Using Parallel Fusion between three Classifiers*, 3rd International Conference on Signals, Circuits and Systems* (2009)*,*1-4

[4] Yong Ma, Chang-chun Bao, Jia Liu, Speaker segmentation and clustering based on the improved spectral clustering, *2011 IEEE International Workshop on Machine Learning for Signal Processing,* (2011), 1-5

[5] Kadri H., Lachiri Z., Ellouze N.: Robustness improvement of Speaker Segmentation techniques based on the Bayesian Information Criterion, *2nd Information and Communication Technologies,* (2006), 1300-1301

[6] Grasic M. and others: The Influence of Speech/Non-speech Segmentation on On-line and Off-line Speaker Segmentation Accuracy, *16th International Conference on Systems, Signals and Image Processing,* (2009), 1-4

[7] Chan W.N. and others: Use of vocal source features in speaker segmentation, *2006 IEEE Conference on Acoustics, Speech and Signal Processing,* (2006), Vol. 1, 657-660

[8] Ziółko B., Manandhar S., Wilson R.C., Phoneme segmentation of speech, *18th International Conference on Pattern Recognition,* (2006), Vol. 4, 282-285

[9] TMS320C6713 DSK technical documentation (2003) http://c6000.spectrumdigital.com/dsk6713/V2/docs/dsk6713_TechRef.pdf

[10] Description of the environment Matlab/Simulink (2012): http://www.mathworks.com/

[11] Stark H. G., Wavelets and Signal Processing, *Springer*, Germany, (2005)

[12] Białasewicz J.T.: Falki i aproksymacje (Wavelets and approximations), *WNT*, Warszawa (2000)

[13] Symlets12 coefficients (2012): http://wavelets.pybytes.com/wavelet/sym12/

***Authors**: prof. dr hab. inż. Adam Dąbrowski,*
dr inż. Paweł Pawłowski, dr inż. Andrzej Meyer,
dr inż. Marek Portalski, mgr inż. Radosław Weychan,
mgr inż. Agata Chmielewska, and mgr inż. Tomasz Janiak
Poznań University of Technology, Chair of Control and System Engineering, Division of Signal Processing and Electronic Systems,
Piotrowo 3a, 60-965 Poznań
E-mail: {Adam.Dabrowski, Pawel.Pawlowski, Andrzej.Meyer; Marek.Portalski, Radosław.Weychan, Agata.Chmielewska, Tomasz.Janiak} @put.poznan.pl