Proceedings of the 5th WSEAS Int. Conf. on MULTIMEDIA, INTERNET AND VIDEO TECHNOLOGIES, Corfu, Greece, August 17-19, 2005 (pp182-185)

Transcription as a Speech Compression Method in Transmultiplexer System

MARIUSZ ZIÓŁKO*, BARTOSZ ZIÓŁKO** and ANDRZEJ DZIECH***
*Department of Electronics,
**Faculty of Electrical Engineering, Automatics, Computer Science and Electronics,
***Department of Telecommunications
AGH University of Science and Technology
al. Mickiewicza 30, 30-059 Kraków
POLAND

Abstract: - A very efficient method of speech compression in telecommunication systems is suggested. It uses the speech recognition system to convert the voice signal into its transcripted form. Next, a speech synthesizer is used to reconstruct speech on the receiver side. Integer filters are used to realize perfect reconstruction in the transmultiplexer system. Although such system destroys the individual speech features, it provides an extremely high compression.

Key-Words: - Transmultiplexing, Compression, Filter Banks, Speech Recognition, Speech Synthesis

1 Introduction

Nowadays there are plenty of telecommunication services. They cannot be dealt with in the same way; however, they should operate in one network, if possible. It is clearly seen that customers have different needs and telecommunications companies do everything to satisfy all of them. For example it has become typical that one mobile telecommunications operator owns two networks: one for business customers (expensive but luxurious) and the other for teenagers and people who prefer cheap solutions without additional services or a great number of offices, etc.

In this paper we present an innovative solution to serve a part of phone calls in a much cheaper way by using an extremely efficient compression, but with loss of information on speaker emotions. An idea of simultaneous transmission of many extremely compressed voice signals by a single channel is presented. A speech recognition system allows the speaker to provide natural, sentence-length patterns. The main task of such systems is to interpret human speech for its transcription.

Transmultiplexation [1] changes the parallel transmission into a serial transmission. The continuous type of processing is very important to process signals in real time.

On the sender side the speech signal is converted into text. Such data need transmultiplexer processing without any distortion. Each bit in each channel must be recovered from the composite signal. On the receiver side text is converted into speech by use of the speech synthesis system.

2 Transmultiplexing

Transmultiplexing is an important method of combining signals of several users into one signal for the transmission by a single channel. Fig.1 shows the classical schematic diagram of a four-channel transmultiplexer. In the transmitter, the *M* input signals were upsampled, filtered and summed to obtain a composite signal. At the receiver end, the composite signal is relayed to four channels of the separation part, where the signal is filtered and downsampled to recover the original input signals. The basic idea [2] is the reversibility of all procedures. A transmultiplexer achieves perfect reconstruction if the output signal s_i^{out} is only a delayed version of the input signal s_i^{in}, namely, if there exists a positive integer τ such that

$$s_i^{\text{out}}(n) = s_i^{\text{in}}(n-\tau), \qquad (1)$$

where *i* is a signal number, $i \in \{1,2,\ldots,M\}$.

There is a growing tendency to equip transmultiplexers in reversible integer-to-integer filter banks [3]. Signals are then processed in finite-precision arithmetic. Due to this property, transmultiplexers of this type can be applied to

transmit lossless compressed signals, low memory is needed and complexity of computations is slight. Such filters are needed in the case when written text is transmitted.

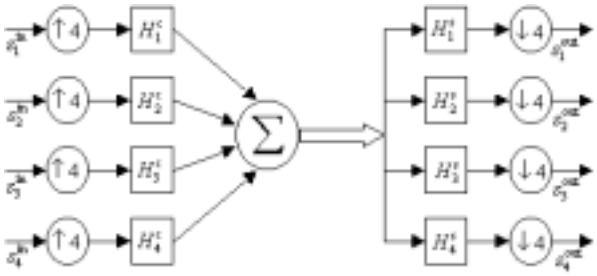

Fig. 1. A scheme of 4-channel transmultiplexer

3 Speech Recognition System

Voiced sounds are produced by modulation of the air flow from the lungs by vibration of vocal cords. Speech is an input in speech recognition system while the sequence of written words is an output. Speech recognition is a problem investigated by many scientists [4], [5], [6].

The time dependent amplitude and the frequency characteristics of a speech signal change in the time domain by continuous reconfiguration of human's voice-tract resonant chambers. There is a number of dynamically altered parameters in the speech signal, which make the analyses and modeling difficult.

Speech signal should be segmented in a certain manner, before analyzing. The signal contained in the obtained segments can be more precisely processed because their characteristics are more constant. The most effective method seems to be constant segmentation for 20 [ms] length blocks. However, good segmentation can be based on energy fluctuations: rises and falls.

The appropriate representation of speech segments is the next important problem. The original time-varying signal representation is not useful in speech recognition. The signal transformation is necessary for an efficient system. Finding an accurate transformation is a fundamental problem. The choice of transform depends on the purpose of analysis. Usually, methods, which are based on the cosine (DCT) or the wavelet (DWT) transform, are used. This way the frequency properties of speech can be analyzed. The analysis of energy distribution for the different frequencies seems to be the best solution to distinguish the phonemes.

Fig.2 presents the waveform for a single word spoken in the Polish language, its time-frequency analysis which is based on DWT and the energy fluctuations.

Each speech recognition system must be equipped with a dictionary which consists of possible words. Especially its size has a great impact on the efficiency of the system. Words in the dictionary should be proper for needs of customers. They should give them the comfort of natural speech in their subjects. However, the dictionary should be as small as possible to increase the speed of processing. Moreover, too many words may cause errors.

Speech recognition consists of a number of elementary steps: voice capturing, noise compensation, segmentation, transformation, feature extraction, classification, lexical correction and grammar correction. Automated speech recognition remains an open problem, since there are only few working and efficient solutions (e.g. Dragon). Speech recognition is useful and demanded by many people. It would considerably increase the speed of cooperation with the computer.

4 Speech Synthesizer

Speech synthesizer is a computer-based system that should be able to read any text aloud, whether it was directly introduced in the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system. It is a much easier problem than speech recognition and there are already very good commercial (i.e. AT&T Labs, Bell Labs) and freeware (i.e. FreeTNT) solutions for many languages. Anyway speech synthesis is still an interesting area for researches. They are focused on modeling emotions for speech synthesis [7]. Speech synthesizers are commonly known as TTS (Text-To-Speech). Synthesizers typically consist of two main blocks. A Natural Language Processing module (NLP) is capable of producing a phonetic transcription of the text read with the desired intonation and rhythm. A Digital Signal Processing module (DSP) transforms the symbolic information it receives into speech.

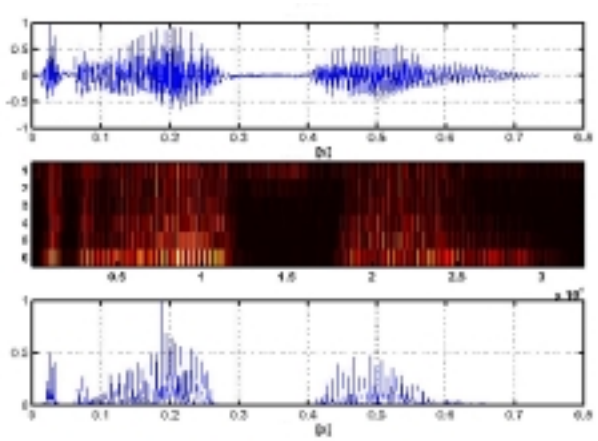

Fig. 2. Word "Rafał" as a signal, its discrete wavelet transform and power waveform

5 Compression

Different types of phone calls demand different reactions. Let us compare two examples. The features of voice are sometimes very important for some speaking people. In that case the exact words are less important. The speech can be noised and even some words can be missed but features of the voice and emotions have to be transferred. Another case is a phone call between employee A and employee B of the same company who do not know each other. The employee A wants to present the production factors of his unit. The emotions and features of voice are not important and sometimes even inadvisable. The task is to transfer digital voice signal without any emotions. In the second case it could be much more effective to send a text instead of voice. For the comfort of employees, a communication system can be equipped with speech recognition and speech synthesis systems on both sides. In that way employees can speak and hear, each other, while their voices are transferred by network as a text.

Bit rate for uncompressed speech signal is 88.2 [kb/s] at 11025 [Hz] sampling rate and 8 bit representation. The compressed bit stream for a signal in a 2-G wireless system is 13 [kb/s] only. It is possible to decrease considerably this bit stream if text is sent to receiver instead of voice. Average phoneme lasts about 50 [ms] in speech waveform. Hoffman compression enables to code written text using 4 bits for one symbol in average. It means that voice presented as text needs a bit stream about 0.08 [kb/s] only. If we compare these results with the classical GSM mobile phone system we obtain the compression ratio 160.

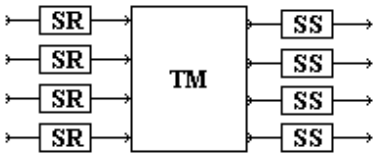

Fig. 3. Voice communication system which is based on the text data format (SR - Speech Recognition, TM - Transmultipler telecommunication system, SS - Speech Synthesis)

Another example is presented in Tab.1. There are two compared representations of the Hamlet monolog. The first one presents the file size where the voice was recorded with 8 bits per sample at 11025 [Hz] sampling rate. The second one is a text representation (transcription). The compression gain for this example is 895!

Table 1. Sizes of text and audio files

File type	**Size [kB]**
Audio (.wav) 8 bit	1203244
Text (.txt)	1344

6 Examples

Let us consider the specific case when filters are FIR type. Let the orders of all combining filters be equal to K which depends on the number of channels in the following way $K \le M-1$. Let the order of all separation filters be $L \le M$ and moreover $h_i^s(0) = 0$ for all $1 \le i \le M$. Let us introduce two matrices

$$G^c = \begin{bmatrix} h_1^c(M-1) & h_1^c(M-2) & \cdots & h_1^c(0) \\ h_2^c(M-1) & h_2^c(M-2) & \cdots & h_2^c(0) \\ \cdots & \cdots & \cdots & \cdots \\ h_M^c(M-1) & h_M^c(M-2) & \cdots & h_M^c(0) \end{bmatrix} \quad (2)$$

$$G^s = \begin{bmatrix} h_1^s(1) & h_2^s(1) & \cdots & h_M^s(1) \\ h_1^s(2) & h_2^s(2) & \cdots & h_M^s(2) \\ \cdots & \cdots & \cdots & \cdots \\ h_1^s(M) & h_2^s(M) & \cdots & h_M^s(M) \end{bmatrix} \quad (3)$$

which consist of combining and separation filter coefficients, respectively. Both matrices are square and their dimensions depend on number of channels. Under these assumptions, the perfect reconstruction conditions (1) can be written in a simple form

$$G^c G^s = E,$$

where E is a unitary matrix. If we assume that both matrices G^c and G^s are nonsingular, then we obtain a simple algorithm for filter designing:

- choose an arbitrary matrix (2) (i.e. coefficients of composition filters)
- compute the coefficients of separation filters

$$G^s = (G^c)^{-1}. \quad (4)$$

It is possible to provide all calculations using the integer numbers only. For this case it is convenient to use such filters that

$$\det G^c = \det G^s = 1.$$

Let us consider an example of a transmultiplexer which consists of four channels and let the following coefficients

$$h_1^c = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix},\ h_2^c = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix},\ h_3^c = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix},\ h_4^c = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad (5)$$

for FIR combining filters be assumed. Simple computations (4) provide the coefficients of separation filters

$$h_1^s = \begin{bmatrix} 0 \\ 1 \\ -1 \\ 0 \\ 0 \end{bmatrix},\ h_2^s = \begin{bmatrix} 0 \\ 0 \\ 1 \\ -1 \\ 0 \end{bmatrix},\ h_3^s = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ -1 \end{bmatrix},\ h_4^s = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}. \quad (6)$$

This example depicts the extraordinary simplicity of filters which can be obtained while using algorithm described in [3].

7 Conclusions

The presented method enables to transmit a large number of speech signals through a single channel. It can be obtained due to strong signal compression. Such system needs to apply the speech recognition and synthesis software. Some disadvantages result from imperfect properties of speech-to-text and text-to-speech converters, especially the recognition system brings noticeable harmful effects. The large variability in the signal makes the speech recognition difficult.

If integer filters are incorporated in filter banks then not only theoretically but also in practice the perfect reconstruction conditions are fulfilled. The usefulness of integer filters enables transmission of text, software files or coded multimedia data like MPEG files.

Acknowledgments

This work was supported by MNiI under grant number 4 T11D 005 23.

References:

[1] M. Vetterli, A theory of multirate filter banks, *IEEE Trans. Accoust. Speech Signal Process*, Vol. 35, 1987, pp. 356-372.

[2] A.N. Akansu, P. Duhamel, X. Lin, and M. Courville, Orthogonal Transmultiplexers, *IEEE Trans. on Signal Processing*, Vol. 46, No. 4, 2001, pp. 979-995.

[3] B.Ziółko, M.Ziółko and M.Nowak, Design of Integer Filters for Transmultiplexer Perfect Recinstruction, *Proceedings of XIII European Signal Processing Conference EUSIPCO-2005*, Antalya. (in appear)

[4] M.Ziółko, M. Kępiński, J. Gałka, Wavelet-Fourier Analysis of Speech Signal, *Proceedings of the Workshop on Multimedia Communications and Services*, Kielce, 2003.

[5] O. Segawa, K. Takeda, F. Itakura, Continuous speech recognition without end-point detection, *Transactions of the Institute of Electrical Engineers of Japan*, Part C, Vol.124-C, No.5, 2004, pp. 1121-7.

[6] O. Farooq, S. Datta, Wavelet based robust sub-band features for phoneme recognition, *IEE Proceedings: Vision, Image & Signal Processing*, Vol.151, No.3, 2004 pp. 187-93.

[7] M. Schroder, Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions, *Affective Dialogue Systems. Tutorial and Research Workshop, ADS 2004*. Proceedings (Lecture Notes in Artificial Intelligence Vol. 3068), Springer-Verlag, 2004, pp. 209-20.