CHANNEL COMPENSATION FOR FORENSIC SPEAKER IDENTIFICATION USING INVERSE PROCESSING

ANDREY BARINOV, SERGEY KOVAL, PAVEL IGNATOV, MIKHAIL STOLBOV

Speech Technology Center (STC), Saint Petersburg, Russia
(barinov, koval, ignatov, stolbov)@speechpro.com

Typically, speaker identification examination requires two audio recordings: a voice sample and a questionable recording. The questionable one is in most of the cases the intercepted or recorded phone call. As mobile phones became the most popular way of communication, the largest number of questionable recordings comes from GSM channels. They use special algorithms and devices to transmit the speech signal through the GSM channel, but these devices and algorithms change the original signal, thus the possibility of usage of such a recording for speaker identification becomes doubtful. In this paper we study how the GSM channel changes the formants position (frequencies) of the speech signal and how inverse filtering helps to compensate for the influence of the channel on speech signal and on forensic speaker identification performance.

INTRODUCTION

In typical forensic speaker identification a law enforcement expert gets one audio recording from the GSM channel and another one from the digital recorder or recording station [1]. Obviously, these two recordings have different quality due to different signal to noise ratio, average frequency response, dynamic range etc [2, 3]. Experts have to work with the recordings which they get and there is no chance to intercept that phone conversation again, but the recording's quality might be low, it is usually necessary to perform channel compensation before identification analysis, which is the subject of the present research. To study how GSM coding influences the biometric traits of a speech signal and how this influence may be compensated with inverse processing, we carried out a series of experiments. In this paper, we analyze only the influence of the channel on the LPC evaluation of formants structure of a speech signal [4]. There are a lot of other types of transmission channels [5] such as analog phone lines, VoIP, satellite etc. but we leave these channels for further researches and in this paper consider only the GSM channel. There are several types of instrumental identification analysis [1- 3], but most of them are based on formants traces analysis. In this paper we ignore all the discussions about the identification technique itself, but focus on the channel compensation and on the improving of the formants representation accuracy.

1 GSM CHANNEL

GSM (Global System for Mobile Communications) is the pan-European cellular mobile standard. Three speech coding algorithms are part of this standard. The purpose of these coders is to compress the speech signal before its transmission, reducing the number of bits needed in its digital representation, while keeping an acceptable quality of the decoded output. As GSM transcoding (the process of coding and decoding) modifies the speech signal, it is likely to have an influence on speaker recognition performance, together with other perturbations introduced by the mobile cellular network (channel errors, background noise). These speech coders, referred to the full rate, half rate and enhanced full rate GSM coders. Their corresponding European telecommunications standards are the GSM 06.10, GSM 06.20 and GSM 06.60, respectively [6]. These coders work on a 13 bit uniform PCM speech input signal, sampled at 8 kHz. The input is processed on a frame-by-frame basis, with a frame size of 20 ms (160 samples).
A brief description of these coders follows.

1.1. Full Rate (FR) speech coder

The FR coder was standardized in 1987. This coder belongs to the class of Regular Pulse Excitation - Long Term Prediction -linear predictive (RPE-LTP) coders. In the encoder part, a frame of 160 speech samples is encoded as a block of 260 bits, leading to a bit rate of 13 kbps. The decoder maps the encoded blocks of 260 bits to output blocks of 160 reconstructed speech samples. The GSM full rate channel supports 22.8 kbps. Thus, the remaining 9.8 kbps are used for error protection. The FR coder is described in GSM 06.10 [6] down to the bit level, enabling its verification by means of a set of digital test sequences which are also given in GSM 06.10. A public domain C-code implementation of this coder is available [7].

1.2. Half Rate (HR) speech coder

The HR coder standard was established to cope with the increasing number of subscribers. This coder is a 5.6 kbps VSELP (Vector Sum Excited Linear Prediction) coder from Motorola [8]. In order to double the capacity

of the GSM cellular system, the half rate channel supports 11.4 kbps. Therefore, 5.8 kbps are used for error protection. The normative GSM 06.06 gives the bit-exact ANSI-C code for this algorithm, while GSM 06.07 gives a set of digital test sequences for compliance verification.

1.3. Enhanced Full Rate (EFR) speech coder

The EFR coder was the latest to be standardized. This coder is intended for utilization in the full rate channel, and it provides a substantial improvement in quality compared to the FR coder [9]. The EFR coder uses 12.2 kbps for speech coding and 10.6 kbps for error protection. The speech coding scheme is based on Algebraic Code Excited Linear Prediction (ACELP). The bit exact ANSI-C code for the EFR coder is given in GSM 06.53 and the verification test sequences are given in GSM 06.54.

So, we can notice that all GSM codecs were especially created for speech signals coding and that they use Linear Predictive model which is the best fit for a speech signal because of its extreme nature [10,11].

2 THE INFLUENCE OF CHANNEL

During transmission of the voice signals through real communication channels, these signals are reproduced with some errors. These errors are usually caused by:
- the distortions from the microphone and channel itself;
-acoustical and electromagnetic interferences and noises affecting the transmitted signal.
The major types of such distortions are frequency, amplitude and phase related. Frequency and time characteristics of the channel define the so-called linear distortion. In addition, the channel can cause nonlinear distortions.
Nonlinearity of the frequency response (FR) of the channel causes the original signal spectrum to change. This effect has crucial meaning for speaker identification performance because it changes the formant's energy and position.
To get the recordings for comparison we conducted an experiment as shown on Fig.1.

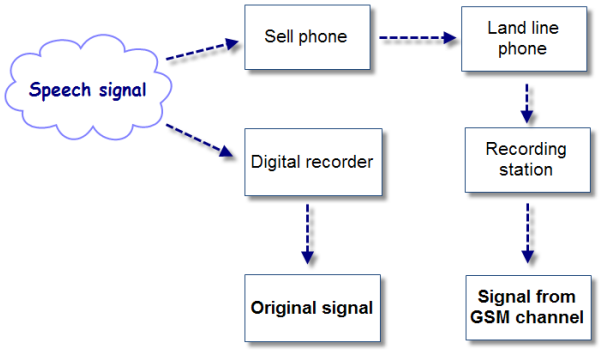

Figure 1: The scheme of experiment.

The speaker made a test call from a cell phone to a landline phone, and this speech signal was recorded in two different manners simultaneously. The first recording (original one) we recorded using a high quality digital recorder, and the second recording (signal from GSM channel) we recorded from the landline phone using a high quality recording station.
In Fig.2 there are two spectra: the average spectrum of original signal and the average spectrum of the signal transmitted through the GSM line. In this case we observe nonlinearity of the GSM channel's frequency response in the range 750-2000Hz which might cause changing in energy distribution and effect 2nd and 3rd formants (F2 and F3). There is also a fall off of the channel's FR at the frequency of 3500Hz which leads to the shifting of the fourth formant (F4).

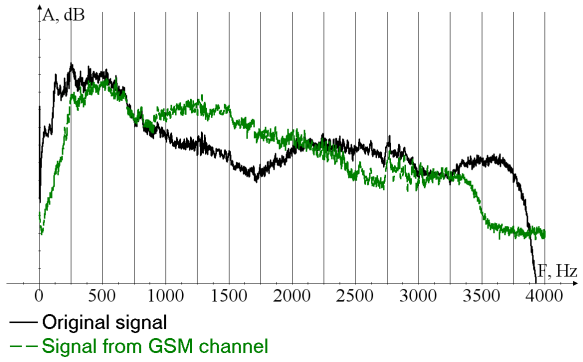

Figure 2: Spectra of the original signal and the signal from the GSM line.

Graphs showing instantaneous LPC spectra of analyzed signals for some vowels are presented in Fig.3-7.

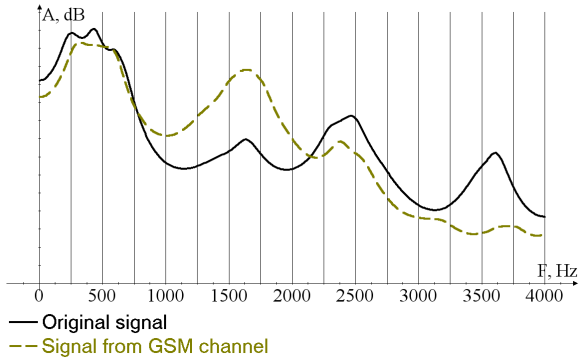

Figure 3: Instantaneous LPC spectra of the original U-like sound and the same signal from the GSM line.

In Fig.3 F2 is reinforced but F4 is shifted and suppressed (in comparison with the original signal).

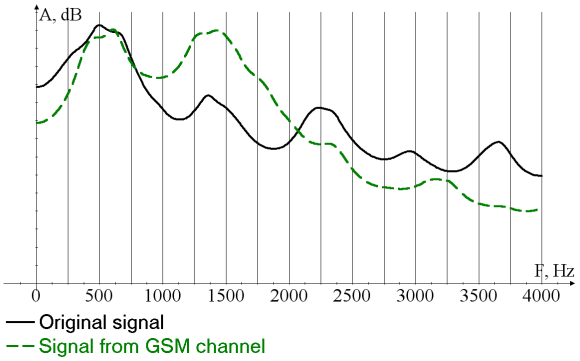

Figure 4: Instantaneous LPC spectra of the original A-like sound and the same signal from the GSM line.

In Fig.4 F2 is reinforced but F3, F4 and F5 are significantly shifted and suppressed.

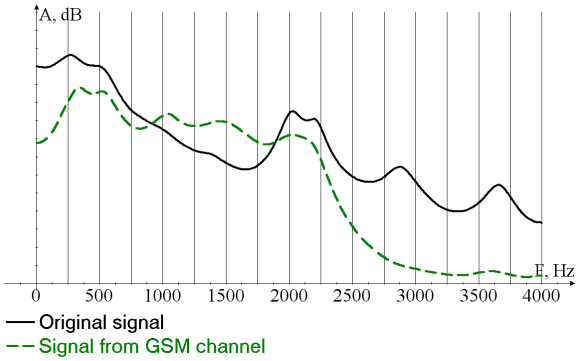

Figure 5: Instantaneous LPC spectra of the original I-like sound and the same signal from the GSM line.

In Fig.5 F3 and F4 are suppressed, at the same time we observe the false maxima in the range 750-2000Hz.

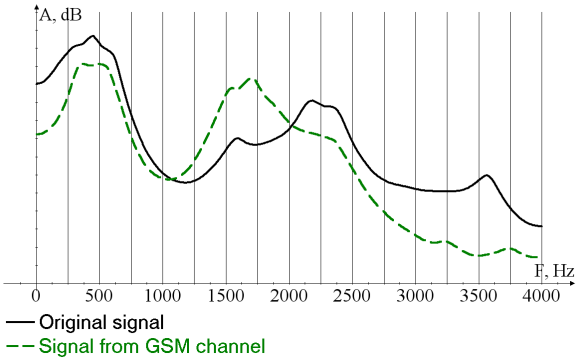

Figure 6: Instantaneous LPC spectra of the original I-like sound and the same signal from the GSM line.

In Fig.6 F2 is reinforced and shifted, F3 is significantly suppressed, F4 is missed but instead of it two false maxima appear.

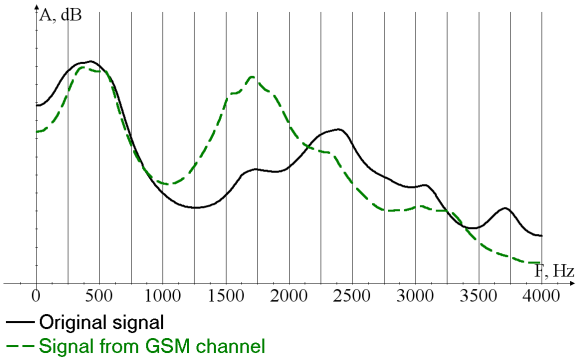

Figure 7: Instantaneous LPC spectra of the original E-like sound and the same signal from the GSM line.

In Fig.7 F2 is reinforced and shifted, F3 is significantly suppressed, F4 and F5 are missing.

3 INVERSE PROCESSING

To compensate for the influence of the transmission channel, considered above, we use inverse processing. Inverse processing can also be used to unmask speech signals and to normalize the frequency response [12,13]. The frequency response of the transmission channel changes the statistical characteristics of the speech signal (spectrum, cepstrum, autocorrelation function). As a consequence, the LPC evaluation of formants may significantly change. To perform the channel compensation, it is necessary to know the transfer function of the channel. The channel is assumed constant for different realizations of the process based on the channel's transfer function estimation. Procedures for compensation in the frequency domain are described below.

In the first procedure, for the recording as a whole or individual segments the average power spectrum is calculated, and then the initial value of the power spectrum of the signal for each frequency is divided into a value of the average power spectrum for a given frequency. Inverse value of the power is the value of the power spectrum obtained by dividing the unit for this value.
The model for signal from the GSM channel is given by:

$$Y(t,f) = S(t,f) \times H(f) \qquad (1)$$

Where Y(t,f) is a short frequency Fourier transformation of the signal transmitted through the channel,

H(f) is a frequency response of the channel, S(t,f) is a short frequency Fourier transformation of the original

speech signal. Using time spectrum averaging for the speech frames, we obtain the following:

$$<Y(t,f)> = <S(t,f)> \times H(f) = S_m(f) \times H(f) \quad (2)$$

Where <…> is a long term averaging operator for the speech frames of the signal, $S_m(f) = <S(t,f)>$ is a model (reference) for the average spectrum of speech signal,

<Y(t, f)> is a long term average spectrum of the signal.

Thus we obtain the equation for channel compensation:

$$\hat{S}(t,f) = Y(t,f) / H(f) = Y(t,f) \times S_m(f) / <Y(t,f)> \quad (3)$$

Where $\hat{S}(t,f)$ is an estimation of the spectrum of the original speech signal, $S_m(f)$ is a model (reference) of the average spectrum of the speech signal.

Because (1) is correct for speech areas, the averaging must also be performed for speech areas of the signal.

In the second procedure, inverse processing can be realized using logarithms of the spectrum. In this case, before performing the calculation of the spectrum and formant frequencies, the spectrum of each phonogram is converted to the form of the logarithm of the spectrum. For the recording as a whole or for its specific parts the average logarithm of spectrum is calculated, and further from the original value of the logarithm of the power spectrum of the processed signal, logarithm of the average power spectrum is subtracted for each frequency.

After logarithmic conversion, equations 1-3 become:

$$Log[Y(t,f)] = Log[S(t,f)] + Log[H(f)] \quad (4)$$

$$<Log[Y(t,f)]> = < Log [S(t,f)]> + Log [H(f)] =$$
$$= LS_m(f) + Log [H(f)] \quad (5)$$

$$Log[\hat{S}(t,f)] = Log[Y(t,f)] - Log[H(f)] =$$
$$= Log[Y(t,f)] + LS_m(f) - < Log[Y(t,f)]> \quad (6)$$

Where LSm (f) is the average logarithm of the speech signal spectrum.

The average spectra are to be calculated for the speech areas of the signal. The detector of tonal speech is used as a speech detector (VAD), because it is the most robust to noise, clicks, tone interference, etc.

Thus inverse processing amplifies the weak average spectrum components and suppresses the strong ones, thus bringing the average spectrum to smooth (flat) one.

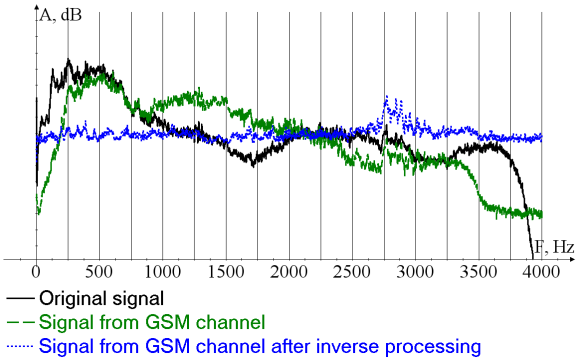

Figure 8: Average spectra of the signals.

Fig.8 shows the average spectra of signals before and after processing. Inverse filtering allows compensation for the nonlinearity of the channel's FR and suppressing false maxima in the spectra. In a more complex, full version of the inverse filtering it is possible to tune the degree of amplifying for the week parts of the spectrum and the degree of the suppression for strong parts of the signal spectrum separately, as well as to define the borders of the processed spectral range for inversion. When the average (channel) spectrum is removed from the current signal spectrum it is reasonable to transform the new, flat spectrum to the standard, typical spectral shape of the input speech signal. There are some examples of inverse processing presented on Fig.9-13.

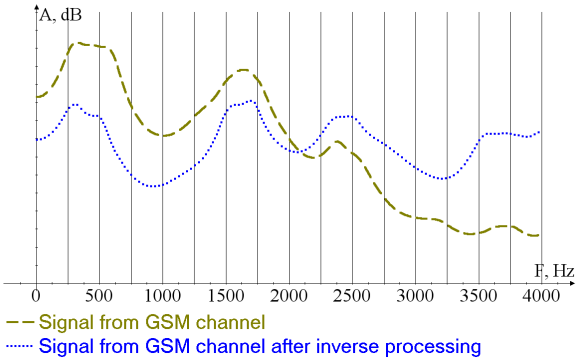

Figure 9: Instantaneous LPC spectra of the signal from the GSM line and after inversion.

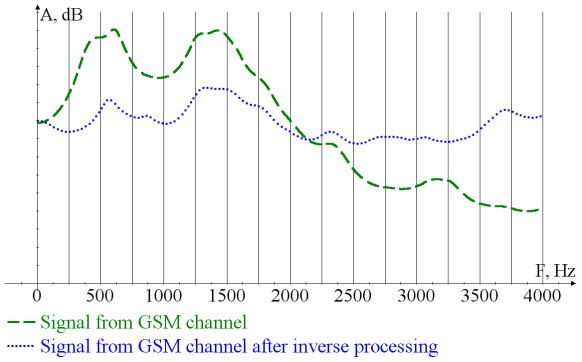

Figure 10: Instantaneous LPC spectra of signal from GSM line and inversed one.

The analysis of the spectra figures reveals that all real (strong) maxima and minima of the spectra remain but some low energy maxima disappear, these are replaced with new maxima which exist in the original signal. The general shape of spectrum after inverse processing becomes more similar to the spectrum of the original signal than the spectrum of the same signal passed through the phone channel.

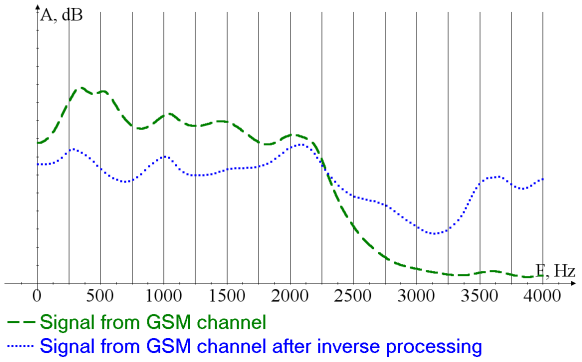

Figure 11: Instantaneous LPC spectra of signal from GSM line and inversed one.

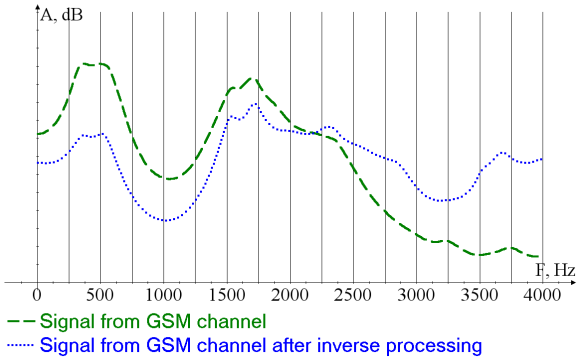

Figure 12: Instantaneous LPC spectra of signal from GSM line and inversed one.

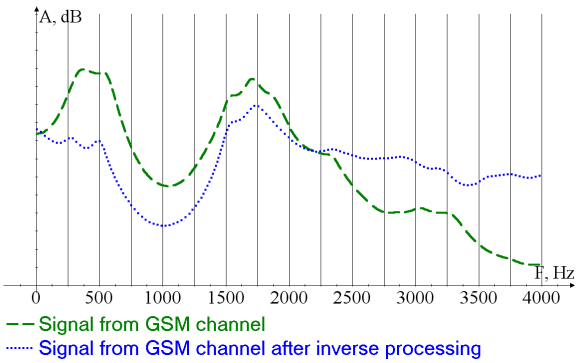

Figure 13: Instantaneous LPC spectra of signal from GSM line and inversed one.

4 RESULTS AND DISCUSSIONS

In this paper we propose to use inverse processing as an effective way to compensate for the influence of the transmission channel on the identification traits of the speech signal.

This processing may be done with two similar algorithms which are described above. In the graphs of Fig.14-18 we present some real-life examples and compare instantaneous LPC spectra of the original signal, the phone line signal and the inversed one. In all cases we have good results and in no cases the process did cause a degradation of results.

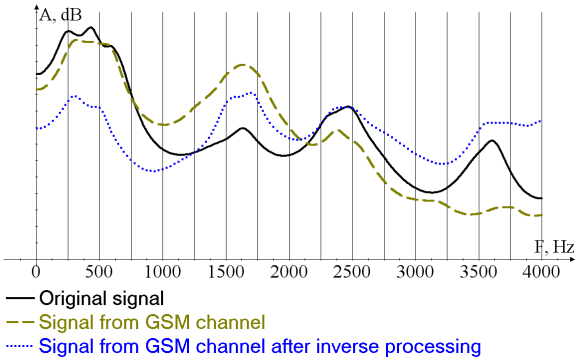

Figure 14: Instantaneous LPC spectra of the original signal, the GSM line signal and the inversed one.

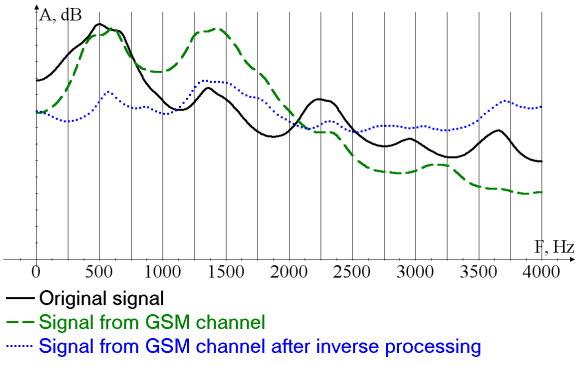

Figure 15: Instantaneous LPC spectra of the original signal, the GSM line signal and the inversed one.

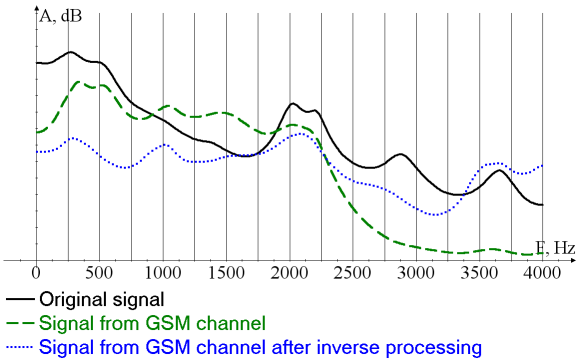

Figure 16: Instantaneous LPC spectra of the original signal, the GSM line signal and the inversed one.

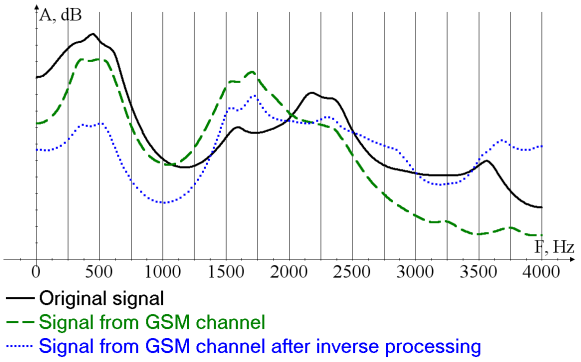

Figure 17: Instantaneous LPC spectra of the original signal, the GSM line signal and the inversed one.

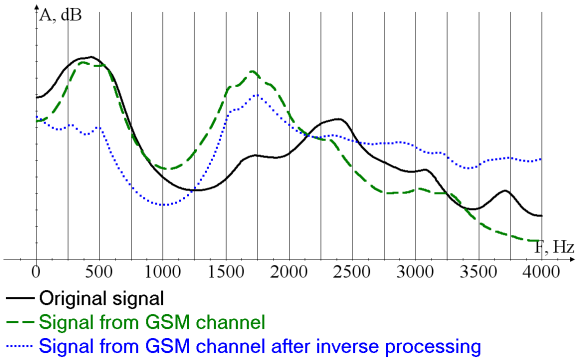

Figure 18: Instantaneous LPC spectra of the original signal, the GSM line signal and the inversed one.

In all graphs above we observe the restoration of the formant structure of the speech signal and it is especially significant in the range of 750-2500Hz (range with nonlinearity of GSM channel's FR) and around 3500Hz where the fall off in GSM channel's spectrum exist.

5 CONCLUSIONS

In this paper we have analyzed the influence of the GSM channel on the formants positions and energy in terms of forensic speaker identification. We also described the methods of channel compensation using inverse processing. The examples presented in this paper show, that the proposed methods of inverse processing allows restoring the original formants structure of a speech signal, which has been corrupted by transmission through low-quality communication channels including GSM lines. The proposed method of channel compensation is more transparent, convenient and effective then well-known methods of cepstral mean substruction, RASTA, etc. It is possible to control in explicit form the spectrum correction in each specific part of the frequency range, for example for the most problematic parts or spectrum - the area of the anty aliasing filter cut-off, in the law-frequencies range bellow 500Hz., and in the spectral areas with law Signal-to-Noise-Ratio. It is possible to choose the best inverse filtering control features manually or automatically with an adaptive self-tuning filter. Moreover, it is experimentally proved that inverse filtration implemented both as a division in spectral domain and with subtracting the corresponding spectral components in the spectral logarithmic domain (method patented by STC) provides a stable positive result.

REFERENCES

[1]Grigoras, C., Cooper, A., Michalek, M. (2009) Forensic Speech and Audio Analysis Working Group - Best Practice Guidelines for ENF Analysis in Forensic Authentication of Digital Evidence, ENFSI – FSAAWG

[2]Laback, B. & W. A. Deutsch (1999): Internal representation of spectral information in dependence of signal type and auditory filter bandwidth. Paper presented at the Joint Meeting 137th Meeting of the Acoust. Soc. Am. and 2nd convention of the European Acoustics Association, Acustica 85, supp 1. 40 (A).

[3]Huber & Runstein, "Modern Recording Techniques", fourth edition, focal press, 1995.

[4] J. D. Markel, A. H. Grey, "linear Prediction of speech", Springer-Verlag, 1976.

[5] C. Byrne and P. Foulkes, "The mobile phone effect on vowel formants", The International Journal of Speech, Language and the Law, 11, pp.83-102, 2004.

[6] hhtp://www.etsi.fr

[7] http:/kbs.cs.tu-berlin.de

[8] I. Gerson and M. Jasiuk, "A 5600 bps VSELP speech coder candidate for half rate GSM", Proc. Eurospeech'93, Vol. 1, pp. 253-256, 1993.

[9] K. Järvinen et al. "GSM Enhanced Full Rate Codec", Proc. ICASSP'97, Vol. 2, pp. 771-774, 1997.

[10] Fant, G. (1960). Acoustic Theory of Speech Production. Mouton & Co, The Hague, Netherlands.

[11] Douglas O'Shaughnessy, "Speech communications Human and Machine", second edition, IEEE press, 2000.

[12]Sound Cleaner. Software for speech signal enhancement and noise cancellation. S Koval, A. Raev, M. Stolbov. Speech Technology Center, St.Petersburg. Russia.1998,http://speechpro.com/eng/products-cleaner.