The invention relates to the system and procedure for hand-free voice communication in video-phone or teleconference using a microphone array, whose main purpose is to make a quality recording of speaker in room, in the situation of larger expansion, with presence noise, with acoustic echo, produced by distance speaker and TV program, room reverberation and movement of the speaker in room. System contains: digital TV receiver and digital camera for picture reproduction and shooting, respectively, stereo loudspeakers and microphone array for sound reproduction and recording, respectively, amplifier and acquisition module for audio signals and DSP for acoustic signal processing. The procedure for microphone signal processing is done in frequency domain and it contains: acoustic echo suppression made of two signals: far-end speaker signal and stereo TV signal, acoustic spatial filtering of near-end speaker in accordance with noise sources and room reverberation, based on adaptive characteristic of microphone array directivity, of speaker localization in horizontal plane, of suppression of all residual noises and adaptive gain control of transmitting signal.
SYSTEM AND PROCEDURE OF FREE SPEECH COMMUNICATION USING A MICROPHONE ARRAY

Technical Field

The invention belongs to the field of acoustic signal processing, precisely speaking to the methods of acoustic echo cancellation, location and selection of an active speaker in the presence of a reverberations in the acoustic environment and the noise suppression by means of microphone array.

Background Art

Hands-free full-duplex speech communication systems are used in many existing applications, such as: video-phone systems, teleconference systems, room and car hands-free systems, human-machine interface using voice, etc.

Usage of the hands-free speech communication systems implies not specified talker position in the acoustic environment, with variable distances from system's microphones and loudspeakers. The hands-free speech communication in such unknown conditions is reason for the number of technical problems, which should be solved, in order to preserve good quality of the speech communication.

Basic problem is acoustic echo generated by partial acoustic energy transmission from a loudspeaker to the microphone, so the speaker on far-end is able to hear his own voice as an obstruction. Conventionally, signal echo canceling is done by adaptive filter using estimation of transfer function of acoustic echo between loudspeaker and microphone, so that its exit gets approximately same signal as acoustic echo signal. Deduction two of these signals cancels acoustic echo. However, canceling echo can not be perfect because of systems non-linearity and acoustics ambience non-steady. As a result it shows residual echo signal. At that basic request stays, recorded speech signal of near-end shouldn't be exposed by echo suppression and its process.

In the acoustic ambient, acoustic disturbances of different nature and causes may appear. Those disturbances could be stationary and non-stationary (for example: computer noises or car noise) and they come from many different sources located on different positions in the room or space where the speaker stands.

Besides that, in closed rooms (as a work rooms, halls and automobile-cabins) it shows up the effect of reverberation as an after effect of multiple acoustic wave reflections from walls and obstacles. Since the acoustic ambient besides the speaker contain sources of disturbances, the desired signal (coming from the speaker) must be separated from the disturbances in order to make possible its own recording. Conventionally, this problem may be solved by using a microphone array having a number of microphones ordered in line at minimum inter-distance. With appropriate processing of microphone array signal, direction dependent sensitivity of microphone system may be achieved. Such microphone systems
has narrow directivity characteristic, enough to record only the actual speaker in the acoustic ambient, while the signals of dislocated noise sources are suppressed, thereby providing higher signal-to-disturbance ratio. The gain depends on: directivity of the microphone array (width of the main lobe), side-lobe size, separability of speech sources and noise sources (to close sources are difficult to separate), reverberation time, non-stationary acoustic sources, etc.

Determination of speaker direction in acoustic ambient and steering the directivity of microphone array according toward is an important problem in hands-free communication systems. The procedures of determining the speaker direction are very sensitive to disturbances present in the ambient, specially: to non-stationary speaker (if it moves within ambient) and if there are several speakers in a given ambient simultaneously speaking (cocktail party effect). The determination of relative direction of the actual speaker to the microphone array in horizontal plane (determination of azimuth), is very important step in video-phone and teleconferencing systems, because of need to determine the speaker coordinates which are used for moveable camera control in the system.

During speech recording in an acoustic ambient, the problem of additive stationary or non-stationary noise always appears so as the residual noise in processing of acoustic signals. They degrade the quality of the recorded speech signal. If they are intense enough, they may even reduce the perspicuity of the speech. There are many algorithms for noise reduction (NR), optimized for specific noise types. The common requirement for all of them is to improve the signal to noise ratio, but to avoid distorting of speech signal and reduction of its perspicuity.

Variable ambient conditions, and variable distance between the speaker and microphone array, require automatic gain control (AGC), which makes the speaker voice level constant and more comfort for the receiver at the far-end of the communication channel. Automatic gain control in full-duplex systems requires additional information from near-end speech activity detector, from far-end speech activity detector and acoustic echo canceller.

Refer to above mentioned technical problems in solution of "hand-free" communication system for speech signal transmission in full-duplex and its usage in video-phone and/or teleconference systems, are very complex. Those problems demand one integral and optimal solution approach, considering real time system operation based on commercial platform of digital signal processor (DSP).

Quality of speech recording in the presence acoustic noises and room reverberations made a complex problem. In the conditions when the useful speech signal spectrum are overlapping with presence noises spectrum, using a single channel processing it is not possible to improve significantly of speech signal quality. In accordance with digital signal processing development and purchasing of enough powerful computer power of DSP, a way of multi-microphone procedure applying acoustic signals processing is open. Benefits of microphone array in relation to single channel processing is adaptation capability of its spatial receipt characteristics (directivity characteristic) to instantly schedule of chosen speaker and define noises in room. At that point, they realize a maximum suppression of presence noises, at the same time the speaker is emphasized. Main problems by microphone arrays usage are (M.S.Brandstein, D.B. Ward (Eds.), Microphone Arrays: Signal Processing Techniques and

When the microphone array is used in video-phone or teleconference systems, in full duplex function, than the number of possible problems is getting larger. The biggest problem is presence of acoustic echo, and then need for automatic gain control (AGC) of system transmitter part, as well as possible presence of system non-steady, called microphony. Additional problem, which is being observed in this patent, is presence of TV program signal, which shows up as an additive acoustic echo on entrance of microphone array.

Large number of mentioned problems has been generated and made very different kind of solutions, which has been patented and which could solve some of problems or few integral problems. For example: U.S. published patent application 2006/ 0153360 Al, filled September 2nd 2005, entitled "Speech signal processing with combined noise reduction and echo compensation", gives integral solution of echo reduction and noise reduction, then U.S. published patent application 7,035,415 B2, filled May 15th 2001, entitled "Method and device for acoustic echo cancellation combined with adaptive beamforming", which gives integral solution of echo reduction and forming of directed microphone array characteristic, then EP published patent application 1 633 121 Al, filled September 3rd 2004, entitled "Speech signal processing with combined adaptive noise reduction and adaptive echo compensation", gives integral solution of residual echo reduction and noise reduction, then EP published patent application 1 571 875 A2, filled February 23rd 2005, entitled "A system and method for beamforming using a microphone array", which gives solution for only directed microphone array characteristic forming, then EP published patent application 1 581 026 Al, filled March 17th 2004, entitled "Method for detecting and reducing noise from a microphone array" gives solution only for noise reduction in microphone array, as well as EP published patent application 1 286 175 A2, filled August 1st 2002, entitled "Robust talker localization in reverberant environment", gives solution only for talker localization in reverberant room.

Integral solution all mentioned problems, realized in this patent, join positive characteristics of particular signal processing of mentioned problems and their solutions, they are going to be solved integrally in frequency domain, optimizing computer resources and gives real time solutions, securing quality of free speech communication in video-phone and/or teleconference systems.

Disclosure of the Invention

Subject of this patent is free speech communication system in video-phone or teleconference applying, which use microphone array and complex acoustic signal processing, which should secure better quality and clearness of speech signal in complex acoustic ambience, in which many previous mentioned failures are separately or integral eliminated.
System, which is subject of this patent, transmits speech and as transmitting medium is being used digital television. For recording and reproduction of speech signal is being used microphone array and loudspeaker, respective, which are integral TV receiver components. When we talk about video-phone or teleconference applying for recording and picture reproduction than we use digital camera and respective digital TV receiver.

Invention essence is specific processing of speech signal, which has been recorded in one acoustic ambience in room where the speaker and system are present. For recording of speaker in room, which stands on define distance (few meters distance) from TV receiver, system uses microphone array of N microphones. Microphone array records all present room signals: useful signal as a directed wave, which gets from the talker to the microphone and different noise signals. As noise signals it shows up: acoustic echo as one loudspeaker direct wave, which is emitting interlocutor voice from the far-end of communication channel, acoustic echo as a directly sound wave, which are emitting stereo TV program, direct waves taken from one or more source of noise or also other sources, which we can hear in the room and reflected waves (room echo), made by their own sources of noise, including speaker, and all those noise, which appear to show during the room reverberation. We should emphasis that noise sources in the room can be stationary or non-steady, which is frequently matter, as by its characteristics, so as by its room location (mobile sound sources).

Different kinds of noises required different techniques for its eliminating, and this invention essence is one optimally designed algorithm, which should at most eliminate all noises and which should secure the best speech signal quality, which is going to be transmitted to the interlocutor on the far-end of communication channel.

Microphone signals from microphone array are being processed in one digital form in DSP, completely in one frequency domain. This domain enables certain advantages, as a processing speed and computer operation number, which is very important for DSP and its real time work. For acoustic echo cancellation it is necessary to put in all loudspeaker signals into the DSP.

DSP run a few complex algorithms: acoustic echo canceling algorithm (AEC), microphone array processing signal algorithm for adaptive beam forming (ABF) and its directivity characteristics, estimation algorithm for direction of arrival (DOA) of useful signal for indoor localization of speaker, in other words speaker room localization, algorithm for reduction of stationary noise, non-steady noise and residual echo (NR- Noise Reduction) and algorithm for system automatic gain control (AGC), because of compensation between different speaker distance from the microphone array. Besides all those basic algorithms, DSP runs some others algorithms more as are: voice activity detector (VAD) on the near-end, VAD on far-end, double talk detector (DTD) on the both sides, additional post filtering (PF) of noise reduction, etc. The aim of mentioned algorithms is maximal reduction of all present noises with minimum of speech signal degradation, therewith secure of transmitting speech signal maximum quality.

Specific aspect of invention subsist adaptive acoustic echo cancellation using an adaptive filter, which mould transferring acoustic way characteristic from loudspeaker to the microphone. Transferring characteristic is complex, working on transmitting way from 2
(stereo) loudspeakers to the N microphone in the microphone array and each microphone signal is being filtered by its on adaptive filter. Work of adaptive filters is being controlled with speech activity detector on the both sides.

Next specific part of invention is adaptive directivity characteristic of microphone array, which secure spatial filtering and directivity separation in the room with speaker, where the useful signal is being boost till the maximum of strength in accordance with and on other signals, which are being interfered. Directivity characteristic of microphone array is accomplished by adaptive weighting and summing of microphone signals, which secure directivity index stability in one frequency domain in one reverberation acoustic ambience.

Defining direction of arrival of speaker directed acoustic wave is a next specific thing of the invention. This system function of free speech communication is necessary for control and managing of directivity characteristic of microphone array by azimuth, also it can be used for control and video camera guiding. It uses microphone signals after acoustics echo cancellation. After generated cross-correlation of microphone signal and its phase transforms, the arrival direction of speakers directed acoustic wave is estimated. This function is being directly controlled by speech activity detector.

Following specific of the invention is process of adaptive suppression of stationary and non-steady noises. Process is realized on the non-linear estimation noise compressor, which is being sorted to several sub-bands. Two estimation noises are being used, securing the optimal suppression result of speech signal characteristics. That has been done because of safety reason. Safety in meaning that process of adaptive noise reduction shouldn't degrade the quality speech signal. Process of filtration should be finished in accordance with adaptive Wiener post-filter.

Specific aspect of the invention is automatic gain control of speech signal before transmission to the far-end interlocutor. This peculiarity is important copulative element of free speech communication system. System secures compensation between different speech signal intensity, as an individual speech characteristic on the one side, and different speech intensity on the other side, which is depending on speaker position, nearer or farther position in relation to the microphone array. The solution makes a difference between speaker activity and useful signal appearing of pause, residual echo, acoustic noise or far-end speech signal, wherefore the solution uses more information previously detected into the system. Analysis of possible scenarios has to be reliable; in counterpart it is possible to get one negative effect of useful speech signal attenuation.

Specialty of this invention is improvement of each mentioned specifics, also improvement in the integration process of all algorithms to the one unite, which functioning is stable and quality. Algorithm procedures are being optimized using cooperative resources.

These and other aspects, specifics and benefits of the invention are going to be more evidently after invention detail description review, patent claims and suitable figures.
**Brief Description of the Drawings**

**Figure 1** - shows elements of free video-phone communication system using a microphone array and digital television.

**Figure 2** - shows ambience conditions for the system appliance of free speech video-phone communication system using a microphone array.

**Figure 3** - shows a diagram block of audio signal processing subsystem within free video-phone communication system; it contains one microphone array with adaptive directivity characteristic (SD-BF), block of speaker indoor location (DOA), block of echo cancellation (AEC), block of noise reduction (NR) and block of automatic gain control (AGC).

**Figure 4** - shows the block diagram of acoustic echo canceling (AEC).

**Figure 5** - shows the block diagram of adaptive determination of near-end speaker direction in horizontal plane (DOA-azimuth).

**Figure 6** - shows the block diagram of spatial filtering (SD-BF).

**Figure 7** - represents the block diagram of noise reduction (NR).

**Figure 8** - represents the block diagram of automatic gain control (AGC).

**Best Mode for Carrying Out of the Invention**

This invention shows a system and method of acoustic signal processing in a free speech communication using a microphone array.

**Figure 1** represents system elements of free video-phone communication using a microphone array and digital television. Digital television 100, which serves the user for a casually TV watching, in the free video-phone communication system, is being used as a video communication and as an audio terminal for audio communication with another speaker. Namely, when the communication channel way 101 gets a call and connection with another speaker is made, then the TV 100 is being used as a multimedia interface, where one speaker over the loudspeakers 102 is listening, and watching on the one part 105 of the TV screen 100 of its far-end interlocutor. In the same time, on the another end of communication channel (far-end side), the speaker on the similar TV receiver, using camera 104 and microphone array 103, also see its interlocutor placed at near-end side. Camera 104 is movable and it is controlled by coordinates, obtained by microphone signal processing from microphone array 103.

Analog signals from a microphone in microphone array 103 are amplified by the amplifier 106 and together with loudspeakers stereo signals 102 are introduced to acquisition module 107, which digitalized them and send them to DSP 108 on the further processing. Proceeded speech signal of the near-end speaker in the DSP is being sent over a communication channel 101 to the speaker on the far-end. Acoustic signal process in DSP 108 gets spatial coordinates of speaker ambience location, in the room with free communication system. With them DSP 108 controls a camera steering 104, directed on the active speaker. On that
way, free audio and video communication between two speakers, with a digital television
system is completely assured.

**Figure 2** schematically shows ambient conditions of free video-phone communication using
a microphone array; it shows only a part of the system, which is related to acoustic signal
processing. The room **201** has installed the system of free video-phone communication,
speaker **202** and noise source **203**, which is normal appearance of every acoustic ambience.
Over the loudspeaker **102** stereo audio system of digital television, the speaker **202** is
listening of incoming speech signal of its interlocutor **204** from the far-end, mostly as a
mono signal. Microphone array (made of N number of microphones) records ambience
sound **201**. After complex microphone signal processing in the block **207**, speech signal of
the speaker **202** is transmitted by the block **208**, to the far-end speaker as a mono signal.

Ambience conditions **201** during the speech communication are very complex, in the case
of the free video-phone communication in the room 201, three noise sources are presence:
stereo loudspeakers **102**, which emit a far-end speaker voice and TV program, speaker **202**
and minimum one source of noise **203**. It is possible that room can have more sources of
noise: computer noise, air-condition noise, street noise, neighbors’ noise, buildings
vibrations or another speaker, or even few speakers, music, etc.

Therefore, we have one very complex acoustic picture of the room. Microphone array **103**
as a sensor system, records all room sounds, and all direct sound waves out of each sound
source, but at the same time, it records all sound reflections. For example, from the
loudspeaker **102** to the microphone array **103** arrives one direct wave **209** followed by
plenty of reflected waves, where only one wave **210** has been showed on the Figure 2, the
speaker **202** sends a direct wave **211** and besides all those waves it sends two more reflected
waves **212a** and **212b**, the noise source **203** sends one direct wave **213** and besides the rest
of waves, one reflected wave **214**, too.

Out of all sounds, which the microphone array records, one is a direct and useful wave **211**
taken from the speaker **202**, all the rest waves are noticed as a disturbances. The biggest
disturbance is an acoustic echo **209**, which comes from the loudspeaker **102**. All other
reflections, together, produce a room reverberation. The task of block for audio signal
processing **207** is to cancel acoustic echo signal, to select a useful signal **211** from the other
signals, to suppress reverberations signals, to suppress direct noise sources and their signals,
and the number of those sources can be more than one. Special task of the **211** block is to
follow acoustic room scene and its non-stationary, depending of speaker mobility, or
position, or depending of noise mobility, are they non-stationary or changeable. In the
following text, explanations of these issues from the invention would be particularly
described.

**Figure 3** shows a schematic diagram of total audio signal processing procedure in free
video-phone communication system using a microphone array. All microphone signals **103**,
from M1 till the M5, as well as a loudspeakers stereo signal **102**, Sp-L I Sp-R, are being
digitalized into acquisition block **107**, **Figure 1**, and converted into the frequency domain
using a fast Fourier transform (FFT) **301** into the signals x/ till the x. It should be
emphasized that the microphone array contains 5 microphones to resolve this patent, but if
there is a need for few additional microphones, they can be install for the need of the
application. The block 302 suppress acoustic echo in all signals (x tilt x 5) using an x 6 and x 7 signals as a refersnts. Suppressed signals SAECS, till S_AECS are being used in the block 304 for assignment of direction of arrival of sound wave (DOA) by horizontal plane (azimuth θ) to the actual speaker. On that way the tracking of the active speaker is possible. Marking the azimuth angle θ in the block 303, the weighted coefficient of signals x, till x 5 are being optimized, with one purpose, to form horizontal directivity characteristic of microphone array with receiving maximum on azimuth direction θ. Receiving characteristic formed in the block 303 has a superdirective nature, which means that the receiving directivity index is larger then directivity characteristic, which we get from delay compensation and sum of microphone signals.

Block 303 does the time compensation between acoustic signal delay of the speaker on the one side, and the microphones on the other side. Control over this delay signal DOA (θ) from the block 304, it is accomplished to control the microphone array directivity by azimuth. Directivity characteristic of microphone array SD-BF (Superdirective Beamformer) in the block 303 is formed. The main lobe of this characteristic is its narrow and directed course, directed into the wanted aim, and the side lobes are intensely slower. That secures spatial filtering to the microphone array, precisely, separation of noise sources in the horizontal plane. That kind of form of directivity characteristic is very important for the reduction of unwanted noises, to separate them from the useful signal and room reverberation effect. Characteristic of directivity has been formed by microphone signal weighting and its summing into the one-channel output signal.

Output signal in block 303 contains constantly speech signal and noise signal, which consists one residual signal after acoustic cancellation of an echo signal, suppressed ambience noise and reduced reverberation noise. That signal comes to the block noise reduction - NR 305 where the additional noise signal reduction is done. Reduction process is adaptive, concerning noise signal non-stationary. Also, important claim in NR realization block is the fact that noise reduction and its process shouldn't affect on speech signal quality.

Final block of signal processing of free speech communication system in video-phone or teleconference processing is block 306 for automatic gain control (AGC) of speech signal. This block uses more information, which it takes out of systems, which are important for defining of possible speech signal conditions and where is necessary to correct its amplitude, on suitable manner. On that way it can be secured almost the same level of transmitting speech signal, independently of the distance between actual speaker and microphone array and it can assure a better quality on opposite side of the communication channel.

On the system exit, the signal process result, using an inverse FFT in the block 307 is transformed from frequency to the time domain. Estimated speech signal on the near-end (S) is sent through the channel to the distant speaker.

Figure 4 represent block diagram of acoustic echo canceling (AEC) 302, which is containing two main blocks: block 401, which is containing 5 adaptive NLMS (Normalized
Least Mean Square) algorithms and block 402, which main function is detection of activities between near-end speaker and far-end speaker speech DTD (Double Talk Detection).

NLMS algorithms, from NLMS1 till NLMS6, processes $x_i$ till $x_f$ microphone signals and certain $s_{AEC1}$ till $s_{AEC5}$ signals to the blocks 303, 304 and 306, Figure 3.

NLMS algorithm function is to cancel echo presence in each microphone signal. This function secures presence of reference signals out of loudspeaker 102 and control signal out of DTD detector 402. NLMS algorithm models transfer functions of acoustic way from each loudspeaker 102 to the each microphone 103: for example, NLMS1 models transfer functions $h_t$ out of loudspeaker Sp-L to the microphone M1 and $u_{RL}$ from loudspeaker Sp-R to the microphone Ml, etc.

Signal transmitted from loudspeaker through NLMS filters, gets a signal replica on the microphone, which came on acoustic way and deduction of these two signals, is accomplished by cancellation of signal echo on the NLMS algorithm exit. To get maximum quality of echo reduction, similar to the case of RLSI type of AEC algorithm (RLS-Recursive Least Squares), it will be described in the text below, DFT coefficient from previous processing blocks are used. NLMS algorithm needs obviously less time in the relation to the RLS algorithm; in the NLMS algorithm realization DTF coefficient of previous 5 processed blocks is being used.

Block 403 with RLSI AEC mark is a main algorithm part of detection procedure of double speech activity from block 402. RLSI AEC does rudely reduction of acoustic noise in the microphone M1 signal using a RLS algorithm. RLS algorithm has a fast convergence, which insures a good estimation of speech signal, as well as an estimation of additive component of signal echo. In accordance with DTF window length of 1024 samples, which is not enough big to secure maximum of noise echo reduction in reverberation room, regression vector gets DTF coefficient out of previous three processed blocks. That process secures double benefit: maximum of echo reduction and signal delays through the system are not enlarged, because of DTF fixed order.

RLSI AEC block exit produce two signals e and y. First signal e is an estimation of near-end speaker voice through the microphone Ml. Second signal $\hat{y}$ is estimation of additive component of echo signal in microphone signal Ml. Both of these two signals are used in detection of double speech activity, which has been realized in the block 402 with DTD mark. Signal from DTD detector controls NLMS algorithm activity, i.e. it stops adaptation of algorithm NLMS 1 to the NLMS 5 algorithm during the double speech activity, when the work of adaptive algorithm is being disturbed. The block 405 does the power averaging of the signal on the loudspeakers by relation:

$$P_{ref} = \frac{|x_0|^2 + |x_f|^2}{2}.$$  \hspace{1cm} (1)

Both signals $\hat{y}$ and $P_{ref}$ are being recursive averaged, in this way we get averaged power of echo signal in the microphone M1 ($z$), and signal in echo generatedolland speakers (3).

$$\overline{P}_{y} = 0.98\overline{P}_{y} + 0.02|j|^2,$$  \hspace{1cm} (2)
\[ P_{\text{ref}} = 0.98 P_{\text{ref}} + 0.02 P_{\text{ref}}. \]  

(3)

Ratio estimation of these two powers is determinate with mark Cs,

\[ C_s = \frac{\sum_{f=0}^{f_2} P_s(f)}{\sum_{f=0}^{f_2} P_{\text{ref}}(f)} \]  

(4)

which is used for power scaling of loudspeakers signal for accomplishing of one soft decision in the block 408. This block determine near-end speaker absence in one microphone signal on the soft decision base, defined with a relation:

\[ D_{\text{id}} = \alpha f \log_{10} \frac{\lambda (C_s P_{\text{ref}}) + \delta}{|s|^2 + \delta}, \]  

(5)

where: \( cc_f \) - is frequency dependent constant, which stiffness favorites allowance of higher frequency convergence, where the signal powers are smaller, however that decrease a possibility of NLMS algorithm divergence. Value \( \lambda \) is the minimum altitude between echo signal power and near-end speaker signal power, whom soft decision is one positive number. Block 409 does limit of control signal \( D_{\text{id}} \) which besides NLMS algorithm leads into the block of DOA-azimuth.

**Figure 5** shows block diagram of the solution for azimuth estimation 304, i.e. determination of the arrival direction of direct sound wave - DOA-azimuth - from an active speaker. Input signals of this block are channel signals \( s_{\text{rec}} \) from AEC block, while output signal is an incoming angle estimation \( \theta_r \). The algorithm is using cross-correlation analysis of the input signals \( S_{\text{aec}} = S_{\text{esc}} \) in block 501, whose outputs represent estimations of the four cross-correlation functions \( G_{1,2}(t,f) = G_{2,1}(t,f) \) using recursive averaging given by:

\[
G_{1,2}(t,f) = \begin{cases} 
\alpha_+ G_{1,2}(t-1,f) + (1-\alpha_+) X_1(t,f) X_2^*(t,f) , & \text{for } |G_{1,2}(t-1,f)| < |X_1(t,f)X_2^*(t,f)| \\
\alpha_- G_{1,2}(t-1,f) + (1-\alpha_-) X_1(t,f) X_2^*(t,f) , & \text{for } |G_{1,2}(t-1,f)| \geq |X_1(t,f)X_2^*(t,f)|.
\end{cases}
\]  

(6)

The constants \( \alpha_+ \) and \( \alpha_- \) should fulfill inequality \( 0.5 < \alpha_+ < \alpha_- < 1 \), with role to increase an influence of the terms \( X_1(t,f)X_2^*(t,f) \) with largest module.

In block 502 marked as PHAT is realized a generalized cross-correlation process known as phase transform. Namely, with usage of the normalized cross-correlation by module, the information about signal energy is lost, while phase information with relative time delay between signals remains. Using inverse FFT transform \( G_{1,2}(t,f) \) and finding its maximum, the assessment of relative time delay between sound waves from two microphones is performed.

Due to formant structure of the speech signal, frequency bins have different power. It is necessary to select frequency bins with highest power and use them to obtain cross-correlation functions. This is why the block 503 performs calculation of the actual power for each channel and power averaging of all signals \( P(t,f) \). In the block 504 is determined filtering function \( W(t,f) \) by emphasizing bins with growing actual signal power. The reason of that is because in the signal segments with abrupt grow of the actual power is main portion of the direct wave, then in segments with declining power dominated by reflected
waves, i.e. room reverberation. In the block 505 is carried out calculation of the average power of the channel signals using both smoothing by frequency and time, \( \bar{P}(t,f) \). The first is performed smoothing of frequency bins by noncausal HR filter of the first order (zero order phase delay is achieved using twofold filtering: forward and backward). Averaging in time is carried out by nonlinear HR filter of the first order with a two averaging coefficients, one involved in power grow and another for a power decline. These nonlinear filters are described with relations:

\[
\bar{P}(t,f) = \begin{cases} 
\alpha_p \bar{P}(t,f) + (1 - \alpha_p) P(t,f) & 0.8 < \alpha_p < \alpha_p < 1. 
\end{cases}
\]

(7)

The variable \( \bar{P}(t,f) \) is used for defining the decision threshold, applied for extraction of the frequency bins with highest power in block 506. Multiplying binary outputs from the block 506 and weighting vector \( W(t,f) \), results in the filtering function \( W(t,f) \), for weighting of the bins of phase transform in block 502. The phase transforms of the cross-correlation functions are additionally filtered in time by HR filter, in order to decrease variance of the correlation function estimations. This describes relation:

\[
\tilde{G}_{k}(t,f) = \alpha_c \tilde{G}_{k-1}(t-1,f) + (1 - \alpha_c) \bar{W}(t,f) \frac{G_{k}(t,f)}{\|G_{k}(t,f)\|}, \quad 0.85 < \alpha_c < 0.95. 
\]

(8)

Besides bin selection using function \( W(t,f) \), it is applied a priori elimination of bins which are out of range of interest. In the block 507 are defined ranges which are a priori out of interest, and its application is before inverse FFT transform (FFT\(^{-1}\)). In block 509 is performed adjustment in time of the cross-correlation functions, then these functions are averaged, its maximum is determined in block 510, whose abcisse represent an estimation of the time delay \( \tau_{est} \). In block 511 a time delay \( \tau_{est} \) is transformed into incoming angle \( \theta_{est} \) of the direct wave from the active speaker.

The estimation of direction of arrival has effect when a speaker is active; otherwise its estimation value from previous active period is taken as current estimation. Detection process of voice activity of the near-end speaker uses information as follows: a) information from block 513 about average power of the microphone signals; b) information from double talk detector \( D_{dt} \) from block 402, Figure 4; and c) information \( \hat{S}_{BF} \) from block 303, the SD-BF, Figure 3. On the basis of this information, in the block 512, a final decision about an activity of the near-end speaker is made. In case that decision about arrival direction is valid, i.e. the near-end speaker is active, a current estimation will be preceded to the output of DOA block 304, otherwise previously valid estimation will be considered as current one.

**Figure** 6 shows block diagram of the forming procedure for the superdirective beamforming filter 303, Figure 3. Due to self-cancellation of the useful signal during application of the adaptive algorithms for canceling acoustic disturbances in reverberant room, instead an adaptive algorithm is often applied superdirective beamforming spatial filter 601 with fixed coefficients. Superdirective beamforming filter supports higher directivity index then a conventional spatial filter using a sum of the delay compensations. Detail description of the forming of the weighting coefficients allowing superdirective characteristics of the filter is given in following text.
The model of the reverberated room is usually in form of diffuse noise field, which consider
present noises from all directions and approximately same intensity. This model of the
diffuse noise field shows that coherence between two microphones is real number equal to:
\[
\Gamma_{ij}(f) = \frac{\sin(2\pi d_{ij} / c)}{2\pi d_{ij} / c},
\]
where \( f \) frequency, \( d_{ij} \) is inter-distance between microphone \( i \) and \( j \), and \( c \) is speed of sound. The coherences of the microphone pair \( F_{ij}(f) \) form coherence matrix \( \Gamma_{ij} \). Using in such way defined coherence matrix \( T_{ij} \), coefficients of the superdirective microphone array are determined in the block 602, according to:
\[
W_{SD}^{H} = \frac{C_{\theta}^{H} H_{\theta} - d}{C_{\theta}^{H} \Gamma_{\theta}^{H} C_{\theta}},
\]
where \( C_{\theta} \) is a vector of directivity in reference to direction of selected speaker defined by an azimuth angle \( \theta \). This vector is determined in the block 603 by:
\[
C_{\theta}^{H} = \left[ 1 \exp\left( -j \omega \frac{d \hat{n}_{j}(\theta)}{c} \right) \cdots \exp\left( -j \omega \frac{4J\sin(\theta)}{c} \right) \right].
\]
The value \( d \) is distance between two neighboring microphones. An output of the block 303 gets a speech estimation \( S_{bf} \) of the actual speaker using equation:
\[
S_{bf} = W_{SD}^{H} C_{AE}.
\]
In Figure 7 the block 305, marked as NR, for noise suppression is shown. Signal \( S_{bf} \) is input signal in the block 305 and it contains both estimated speech signal and signal residuals from disturbances originated from an acoustic echo, an acoustic indoor interferences and a room reverberation. Signal \( S_{bf} \) is entering in the block 701, marked as FWF⁻¹, where IFFT is executed, then additional windowing of the signal segments in time scale is carried out in order to make 'soft' cutting of the segment's ends, and finally, return back to the frequency domain using FFT. An essence of this operation is as follows: during previous signal processing steps, an equivalent time signal form is extended to the FFT window ends. Using new operations of the Wiener filter, additional expansion of the segments and cyclic overlapping of their ends is performed, which makes impulse interferences, heard as monotonous "crackling". Applied FWF⁻¹ procedure is completely removing above mentioned issue, without introducing any additional signal distortion.

In next two blocks 702 and 703, a noise estimation using minimum power of the input signal is performed. Hence an adaptation using actual minimum power is not satisfactory, due to DFT coefficients of the certain segments with extremely low power could disturb previous estimation of the noise power, the noise estimation is realized using three processing blocks: the first block 702 carries 'out slow estimation of the noise power \( \tilde{N}_{dolh} \), the second 703 performs fast estimation of the noise power \( \tilde{N}_{fsl} \), and the third block 704 executes actual estimation of the noise power \( \tilde{N} \) using nonlinear transform of the both \( \tilde{N}_{aelh} \) and \( \tilde{N}_{fsl} \) estimates.
The fast and the slow estimates of the noise power are realized using same recursive moving average HR filter of the first order with different adaptation factors for grow and decline of the output value:

\[
\hat{N}_{\text{slow}}(t) = \begin{cases} 
\alpha_{\text{slow}} \hat{N}_{\text{slow}}(t-1) + (1-\alpha_{\text{slow}}) |S_{\text{SBP}}(t)|^2, & \text{for } \hat{N}_{\text{slow}}(t-1) < |S_{\text{SBP}}(t)|^2, \\
\alpha_{\text{slow}} - \hat{N}_{\text{slow}}(t-1) + (1-\alpha_{\text{slow}}) |S_{\text{SBP}}(t)|^2, & \text{for } \hat{N}_{\text{slow}}(t-1) \geq |S_{\text{SBP}}(t)|^2, 
\end{cases}
\]

\[
\hat{N}_{\text{fast}}(t) = \begin{cases} 
\alpha_{\text{fast}} \hat{N}_{\text{fast}}(t-1) + (1-\alpha_{\text{fast}}) |S_{\text{SBP}}(t)|^2, & \text{for } \hat{N}_{\text{fast}}(t-1) < |S_{\text{SBP}}(t)|^2, \\
\alpha_{\text{fast}} - \hat{N}_{\text{fast}}(t-1) + (1-\alpha_{\text{fast}}) |S_{\text{SBP}}(t)|^2, & \text{for } \hat{N}_{\text{fast}}(t-1) \geq |S_{\text{SBP}}(t)|^2, 
\end{cases}
\]

wherein the constants \( \alpha_{\text{fast}}, \alpha_{\text{slow}}, \beta_{\text{fast}}, \beta_{\text{slow}} \) are related by:

\[
0.2 < \beta_{\text{fast}} < \beta_{\text{slow}} < \alpha_{\text{fast}} < \alpha_{\text{slow}} < 1 >
\]

The fast and the slow noise estimates are combined in block 704, marked as nonlinear compressor. The final noise estimation is given by relation:

\[
\hat{N} = \begin{cases} 
\beta \hat{N}_{\text{slow}} \left( \frac{\hat{N}_{\text{fast}}}{\hat{N}_{\text{slow}}} \right)^{\alpha}, & \text{for } \hat{N}_{\text{fast}} > \hat{N}_{\text{slow}}, \\
\beta \hat{N}_{\text{fast}}, & \text{for } \hat{N}_{\text{fast}} \leq \hat{N}_{\text{slow}}, 
\end{cases}
\]

where parameter \( \alpha (0.25<\alpha<0.5) \) is controlling compression level of the noise estimation dynamics, while parameter \( \beta \) is defining overestimation of the noise power. The meaning of the nonlinear transform is as follows: in case of \( \hat{N}_{\text{fast}} > \hat{N}_{\text{slow}} \) usage of the fast estimation only, will result with excessive suppression of the speech signal as well, hence compression of the noise estimation dynamics is introduced. In case when \( \hat{N}_{\text{fast}} \leq \hat{N}_{\text{slow}} \), the compression is not applied in order that faster declining of the noise estimation. Hereby the cutting of phoneme's part at word endings is disabled, when due to fast declining of a signal power, previous value of the noise estimate from the slow estimator, could not follow this dynamic changes. Because of unfavorable speech signal - to - noise ratios at high frequencies, it is defined set of parameters \( \alpha \) and \( \beta \) for 4 characteristic frequency bands (0-2000Hz), (2000-2500Hz), (2500-3500Hz) and (3500-5012Hz), according to an expected signal-to-noise ratio. This parameter set is stored in block 705.

In block 706 is carried out Wiener filtering using transfer function:

\[
h_w = \sqrt{\max \left\{ \frac{|e_i|^2 - \beta_{e_2} \hat{N}}{|e_i|^2}, 0 \right\}},
\]

where the constant \( \beta_{e_2} \) has an estimation function, which should achieve balance between higher noise suppression rate and minimum degradation of the useful speech signal as an initial assessment of the noise power. The transfer function \( h_w \) could have unacceptable long impulse response in time domain, which is producing degradation at DFT block ends, hereby is introduced "soft" cutting of the impulse response using above described FWF\(^3\) procedure. At the end of block 707, additional filtering of the output estimated speech signal \( \mathcal{S} \), carries out, in order to remove spectral components outside of speech range, which could affect doing of AGC block.

**Figure 8** shows the automatic gain control block (AGC) of the system output signal, block 306. AGC task is: (1) to boost the weak speech signals, and to make weaker to strong
signals in accordance with previously determined characteristic of signal dynamics compression, (2) on the input signal parts, where the only echo signal is present, stationary noise or concurrent speaker - noise, and to allay these noises, and (3) to allay input signal parts, where both signals are present, a useful and disturbance signal, and to kept speech clearness.

On the block entrance 306 arrives $s_{ny}$ signal out of block NR, Figure 3, blocks 305, and gets through the compressor of the signal dynamics with adaptive slope characteristic of compression, block 801. Block 801 exit is a signal $s_{AGC}$ - which goes to the block 307, Figure 3, where the inverse Fourier transform FFT converts out of frequency domain into the time domain, as a final estimation signal of speech signal $\hat{s}$, transferred to far-end speaker through the digital television channel.

Automatic gain control is done in the block 801 in accordance with next relation:

$$A_{\text{agc}} = \left( \frac{P_{\text{nom}}}{P_{lu} + \alpha P_{\text{nom}}} \right)^{(1 - \text{SLOPE})/2},$$

(18)

where the: $A_{\text{agc}}$ - as a gain of AGC block, $P_{\text{nom}}$ - nominal power of the output signal, $\alpha$ - constant of limiting maximum gain on the level $A_{\text{agc max}} = -\text{dB}$ (for value $\alpha=0.001$, the maximum gain is $A_{\text{agc max}} = 31.6$ dB), $P_{lu} = P_d + P_n + P_{\text{echo}}$ ($P_d$ - the power of the useful speech signal, $P_n$ - power of diffuse ambience noise, and $P_{\text{echo}}$ - signal power of residual echo), and $\text{SLOPE} = \ln [P_{dp}(t)]$ - which represents a compression level of signal dynamics in the form of complex function of the peak power of useful speech signal. The block 802 calculates a SLOPE, based on the trajectory analysis of the useful speech signal peak power and based on pursuing of its convexity and growing trend.

The block 803 calculates a useful speech signal peak power due the following relations:

$$P_{dp}(t) = \begin{cases} P_d(t), & \text{if } P_d(t) \geq P_{dp}(t-1) \\ \alpha_d P_{dp}(t-1) + (1 - \alpha_d) P_d(t), & \text{if } P_d(t) < P_{dp}(t-1) \end{cases},$$

(19)

where the $\alpha_d$ - is a near constant of the value 1.

The block 804 defines an estimation power of residual echo according to the relation:

$$P_{\text{echo}} = \alpha_{\text{echo}} |\hat{y}|^2,$$

(20)

where $\alpha_{\text{echo}}$ is a constant of echo signal cancellation $\hat{y}$ from block 402, Figure 4.

The block 805 does diffuse noise estimation of $P_n$ as a difference of mean power value of input signals $s_{\text{AEC1}}$ till $s_{\text{AEC3}}$ into the block 303, Figure 3, and a output signal $s_{\text{BF}}$ power out of block 303.

Inline relation procedure for $A_{\text{agc}}$ for forward assigned value of SLOPE, doesn't give positive results; it treats the rest of the noises and useful signal in same way. If we have only a noise presence, than they enhance, what is not good. Therefore, we have to detect and separate following cases: (a) pause in the useful speech signal, (b) presence of the residual
echo, and (c) presence of concurrent speaker or acoustics disturbance. When one of these cases is detect, variable SLOPE is equal to 1 and that stops noise enhancing.

Useful speech signal pause has a different stationary than speech signal. Speech signal, even a week one, is timely non-stationary, while speech pause has a presence of one slowly changing ambience noise. Linear trend of the normalized signal power would be a good indication of signal non-stationary. Furthermore, the trajectory convexity indication should be negative on the local maximum.

**Industrial Applicability**

This invention describes methods of an acoustic and speech signal processing in full-duplex free speech communication system. It relies on free speech communication in one digital television system, but at the same time it can be used for others communication systems, as are video-phone systems, teleconference systems, speakerphones in the room or car, human-computer voice communication, etc. A specific solution found in this invention is its integration into the standard digital TV receiver and its optimization for indoor environment with middle reverberation value till the 600 ms.

Techniques and procedures of acoustic and speech signal processing analyzed in this invention can be generalized on N microphones in the microphone array by the multi-channel recording and on M loudspeakers by multi-channel reproduction.

Techniques and procedures of the acoustic and speech signal processing, analyzed in this invention are under control of larger parameters number, which realize solution optimization of different kind of application.

Methods and techniques of acoustic and speech signal processing, analyzed in this invention could be implemented on the more different ways. For example, these techniques can be implemented into the hardware, software, or in the combination two of them. In the hardware implementation we can use specific integrated circuits (ASIC), processors of the digital signal processing (DSP), programmable logical devices (PDL or FPGA) and others electronic circuits, designed in that way, to be able to accomplish a given invention functions.

Methods and techniques of acoustic and speech signal processing can be implemented into the software or into the software modules. Program codes can be memorized into the memory unit and can work with processors such as a PC, PDA, DSP, etc.

The invention details here described support and help every technician skilled in the relevant art, to implement these principles into another system of free speech communication, without departing from scope of this invention.
Claims

1. The system for a hands-free speech communication using a microphone array, which contains a digital TV receiver that allows audio and video communication facilities in full duplex wherein the digital TV receiver (100) performs a stereo audio reproduction (102) of the stereo TV signal and a mono reproduction of an incoming speech signal needed for a video-telephone communication, and which has a moving video camera (104) for a speaker's recording in a room and presenting a picture of the remote speaker on a window of its screen (105); which contains a microphone system embedded (103) in the TV receiver (100) that records the voice of the speaker and other surrounding sounds at the near end, and that has the purpose to locate the position of the speaker in the room and to control the direction of the video camera (104).

2. The system according to claim 1, wherein the audio transmitter (207) and (208) allows the suppression of the local acoustic echo (209) that is generated in the loudspeakers of the TV receiver (102), suppress the surrounding noise (213) and reverberations (210), (212), (214), determines the location of the speaker in the room, performs an adaptive control of the transmitted signal level and gives the coordinates that are used to control the video camera.

3. The system according to claim 2, wherein it contains a microphone array (103) composed of more than 2 microphones that produce microphone signals for further parallel processing, a module for adaptive cancellation of the acoustic echo (AEC) (302) that is composed of a set of adaptive filters, a module for the direction of arrival (DOA) estimation of the speaker's direct sound wave (304) and beamforming control of the microphone array, a module that forms superdirective beamforming characteristics of the microphone array that has the optimized ratio between the peak of mainlobe level and the peak sidelobe level (SD_BF) (203), a module for adaptive reduction of the surrounding noise (NR) (305), and a module for automatic gain control (AGC) (306).

4. The system according to claim 3, wherein it contains a microphone array (103) with equidistant microphones located in the horizontal plane and mounted along the upper edge of the TV receiver (100).

5. The systems wherein it cancels acoustic echo (209) that is generated in stereo loudspeakers (102) of the TV set and is composed of both a stereo audio TV signal (205) and a mono speech signal that originates from a far-end speaker (204).

6. The system according to claim 5, wherein the unit for echo cancellation (302) and the unit for suppression of surrounding noise (305) work also under conditions of low signal to noise ratio.

7. The system according to any of the previous claims wherein it allows an adaptive location and azimuth tracking of the moving speaker in the room.
8. The system according to claim 7 wherein it allows an adaptive determination of the azimuth coordinate needed for the video camera control.

9. The system according to claim 4 wherein its microphone array forms a narrow beamform directivity that allows the spatial filtering and separation of the actual speaker from the other sound sources in the room.

10. The system according to claim 9 wherein its microphone array forms a narrow beamform directivity which allows the echo cancellation in the room that is produced by the sound wave reflections or by the room reverberations.

11. The system according to any of the previous claims wherein it maintains, by using AGC of the system, an average level of the transmitted speech signal in acceptable limits of the normal voice signal dynamic that does not depend on the distance and the position of the speaker in respect to the microphone array.

12. The technique for hands-free full duplex speech communication using microphone arrays, wherein it performs parallel processing of microphone signals generated in the microphone array and thus adaptively cancels acoustic echo in the microphone signals, performs the direction of arrival estimation of the direct sound wave of the near-end speaker, forms a superdirective beamforming characteristic of the microphone array and controls its azimuth coordinate, suppresses all noise signals contained in the microphone signals and performs an automatic control of the level of the transmitted voice signal.

13. The technique according to claim 12, wherein the complete processing of all audio signals is done in frequency domain.

14. The technique according to claim 12, wherein the adaptive cancellation of the acoustic echo is done for each microphone signal separately and the cancellation includes both signals that come from the stereo loudspeakers.

15. The technique according to claim 14, wherein the adaptive suppression of the acoustic echo is done for each microphone signal separately by means of normalized least mean square (NLMS) algorithms (401) that are controlled by the detectors of the voice activity at both sides (double talk detectors - DTD) (402).

16. The technique according to claim 14, wherein the NLMS algorithms are controlled by means of a detector of voice activity at the near-end that is designed inside DTD and based on the recursive least square (RLS) (403) adaptive algorithm under special conditions that are defined by the continuous presence of the TV audio signal that contains the speech signal alongside with (background) music/speech signal.

17. The technique according to claim 12, wherein the DOA estimation of the direct sound wave of the actual speaker is based on the cross-correlation analysis of the microphone signals after suppression of the acoustic echo.
18. The technique according to claim 17, wherein the DOA estimation of the direct sound wave of the actual speaker is controlled by a voice activity detector (VAD) for speech signal at the near-end.

19. The technique according to claim 12, wherein the directional characteristic of the microphone array is formed in the SD-CBF (303) module as a superdirective beamforming characteristic that is based on the principle of weighting and summation of microphone signals after the completion of acoustic echo suppression and adaptive azimuth control.

20. The technique according to claim 12, wherein the coefficients of the superdirective beamforming microphone array are determined by means of a coherence function of the microphone signals and a directivity vector, where the directivity is taken in respect to the direction of the selected speaker defined by the azimuth angle.

21. The technique according to claim 12, wherein the function of the surround noise suppression is achieved by means of adaptive Wiener filter.

22. The technique according to claim 21, wherein the estimation of residual noise in the noise suppressor is optimized according to the characteristics of the voice signal and realized as a nonlinear compressor of the dynamics of the estimated noise which is frequency dependant.

23. The technique according to claims 12 and 22, wherein the module for automatic gain control of the system is based on the compressor of signal dynamic that has an adaptive slope of compression characteristics (801).

24. The technique according to claim 23, wherein the compressor of the voice signal dynamic is controlled by a detector that indicates the presence of residual acoustic echo, a detector of the pause in the speech signal, and a detector of a concurrent speech and acoustic noise.
FIG. 3.

FIG. 4.
**FIG. 5.**

**FIG. 6.**
FIG. 7.

FIG. 8.