

(19) Japan Patent Office (JP)

(12) **Japanese Unexamined Patent
Application Publication (A)**

(11) Japanese Unexamined Patent
Application Publication Number

**Japanese Unexamined Patent
Application Publication No.
H11-305792**

(43) Publication date: November 5, 1999

(51) Int. Cl. ⁶	Identification codes	FI	Technical indications
G 1 0 L 3/02 9/00	3 0 1	G 1 0 L 3/02 9/00	3 0 1 E F
Request for examination: Not yet requested Number of claims: 37 OL (Total of 22 pages)			
(21) Application number	Japanese Patent Application No. H10- 232565	(71) Applicant	000004226 Nippon Telegraph & Telephone Corporation 2-3-1 Otemachi, Chiyoda-ku, Tokyo
(22) Date of application	August 19, 1998	(72) Inventor	Tomohiro TAKANO % Nippon Telegraph & Telephone Corporation, 3- 19-2 Nishishinjuku, Shinjuku-ku, Tokyo
(31) Priority claim number	Japanese Patent Application No. H10- 39206	(72) Inventor	Hiroyuki MATSUI % Nippon Telegraph & Telephone Corporation, 3- 19-2 Nishishinjuku, Shinjuku-ku, Tokyo
(32) Priority date	February 20, 1998	(74) Agent	Patent Attorney Takashi KUSANO (and one other)
(33) Priority country	Japan (JP)		

(54) (TITLE OF THE INVENTION) SOUND COLLECTION DEVICE, VOICE RECOGNITION DEVICE, METHODS THEREOF, AND PROGRAM RECORDING MEDIUM

(57) (ABSTRACT)

(PROBLEM) To accurately estimate noise, to enable the suppression of non-stationary noise, and to improve a voice recognition rate.

(SOLUTION) Outputs of a microphone 1 which is close to a target signal sound source and a microphone 2 which is farther away are respectively converted into spectra (3, 4); a level difference $\Delta LR(\omega_i)$ between respective corresponding components $L(\omega_i)$ and $R(\omega_i)$ is detected (5); a (ω_i) for which $\Delta LR(\omega_i)$ is not less than a threshold value $Th(\omega_i)$ is selected (6); a component $L(\omega_i)$ is extracted from $L(\omega)$ and converted to a time waveform (8); a component $L(\omega_k)$ for which $\Delta LR(\omega_i)$ is not greater than $Th(\omega_i)$ is given a weight $w(\omega_k)$ (9) and converted to a time waveform, and a sound source silent section is determined (10); if $\Delta LR(\omega) > Th(\omega)$ at this time, then $\Delta LR(\omega)$ is set as $Th(\omega)$ (11). The converted time waveforms are accumulated, and voiced sections are detected with a bone conduction microphone. Voice recognition is performed for accumulated waveforms only in those sections.

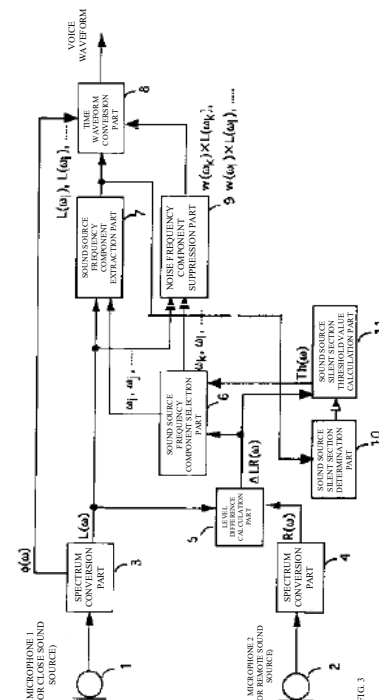


FIG. 3

(SCOPE OF THE PATENT CLAIMS)

(CLAIM 1) A sound collection device comprising:

a first microphone installed at a position close to a sound source of a target signal;

a second microphone installed at a position farther from the sound source of the target signal than said position;

first and second spectrum conversion means for converting respective output signals of the first and second microphones to amplitude spectra and phase spectra;

a level different calculation means for calculating a level difference in amplitude spectra for each corresponding frequency component outputted from the first and second spectrum conversion means;

a sound source frequency component selection means for comparing the level difference for each frequency component outputted from the level difference calculation means and a preset threshold value, and determining a frequency component for which the level difference is greater than the threshold value to be a frequency component for which the target signal is dominant;

a sound source frequency component extraction means for extracting a frequency component for which it is determined that the target signal is dominant based on said determination from an amplitude spectrum of an output signal of the first microphone; and

a time waveform conversion means for converting the amplitude spectrum extracted from the sound source frequency component extraction means to a time waveform using the phase spectrum calculated by the first spectrum conversion means.

(CLAIM 2) The sound collection device according to Claim 1, comprising:

a noise frequency component suppression means for applying a preset weighting factor to an output frequency component of the first spectrum conversion means associated with a frequency component for which it is not determined that the target signal is dominant in the sound source frequency component selection means;

wherein in the time waveform conversion means, the frequency components processed by the sound source frequency component extraction means and the noise frequency component suppression means are converted to time waveforms using the phase spectra of the corresponding frequencies calculated by the spectrum conversion means.

(CLAIM 3) The sound collection device according to Claim 1 or 2, comprising:

a sound source silent section determination means for comparing a magnitude of an output amplitude spectrum of the sound source frequency component extraction means and a preset silent section determination threshold value and determining that the sound source of the target signal is in a silent state when the amplitude spectrum is smaller than the silent section determination threshold value; and

a threshold value calculation means which, when it is determined by the sound source silent section determination means that the sound source of the target signal is in a silent state, updates the threshold value used in the sound source frequency component selection means to at least the level difference outputted from the level different calculation means.

(CLAIM 4) The sound collection device according to Claim 2 or 3, comprising:

a weighting factor calculation means for calculating a weighting factor to be applied to a frequency component for which it is not determined that the target signal is dominant in the noise frequency component suppression means by comparing a magnitude correlation of the level difference of each frequency component from the level different calculation means and the threshold value of the sound source frequency component selection means.

(CLAIM 5) The sound collection device according to Claim 3 or 4, comprising:

a sound source silent section attenuation means for attenuating the outputs of the sound source frequency component extraction means and the noise frequency component suppression means or the output of the time waveform conversion means when it is determined by the sound source silent section determination means that the sound source of the target signal is in a silent state.

(CLAIM 6) The sound collection device according to Claim 1, 2, 3, 4, or 5;

wherein the first microphone and the second microphone are incorporated into a handset, a headset, an ear microphone set, or the like so that the latter is situated at a position closer to the mouth of a user than the latter during use.

(CLAIM 7) A sound collection method comprising:

installing a first microphone at a position close to a sound source of a target signal;

installing a second microphone at a position farther from the sound source of the target signal than said position;

converting respective output signals from the first and second microphones to first and second amplitude spectra and first and second phase spectra;

calculating a level difference in the first and second amplitude spectra for each corresponding frequency component;

comparing the calculated level difference for each frequency component and a preset threshold value and determining a frequency component for which the level difference is greater than the threshold value to be a frequency component for which the target signal is dominant;

extracting a frequency component for which it is determined that the target signal is dominant based on said determination from the first amplitude spectrum based on the determination; and

converting the extracted frequency component to a time waveform using the first phase spectrum.

(CLAIM 8) The sound collection method according to Claim 7; wherein a preset weighting factor is applied to the first amplitude spectrum of a frequency component for which it is not determined that the target signal is dominant in the comparison with the threshold value; and

at the time of the time waveform conversion, the frequency component to which the weighting factor is applied is also converted to a time waveform using the first phase spectrum of the corresponding frequency.

(CLAIM 9) The sound collection method according to Claim 7 or 8; wherein a magnitude of a first amplitude spectrum of a frequency component for which it is determined that the target signal is dominant is compared to a preset silent section determination threshold value, and it is determined that the sound source of the target signal is in a silent state when the first amplitude spectrum is smaller than the silent section determination threshold value; and

when it is determined that the sound source of the target signal is in a silent state, the threshold value used to determine whether the frequency component is one in which the target signal is dominant is updated to at least the level difference.

(CLAIM 10) The sound collection method according to Claim 8 or 9;

wherein a weighting factor to be applied to a frequency component for which it is not determined that the target signal is dominant is calculated by comparing a magnitude correlation of the level difference of each frequency component and the threshold value.

(CLAIM 11) The sound collection method according to Claim 9 or 10;

wherein when it is determined that the sound source of the target signal is in a silent state, a frequency component which is extracted after it is determined that the target signal is dominant and a frequency component for which it is not determined that the target signal is dominant and to which the weighting factor is applied, or the signal converted to a time waveform, is attenuated.

(CLAIM 12) The sound collection method according to Claim 7, 8, 9, 10, or 11;

wherein the first microphone and the second microphone are incorporated into a handset, a headset, an ear microphone set, or the like so that the latter is situated at a position closer to the mouth of a user than the latter during use.

(CLAIM 13) A recording medium for recording a program for making a computer of a sound collection device execute the following processes:

a process for converting respective output signals of a first microphone installed at a position close to a sound source of a target signal and a second microphone installed at a position farther from the sound source of the target signal than said position to first and second amplitude spectra and first and second phase spectra;

a process for calculating a level difference in the first and second amplitude spectra for each corresponding frequency component;

a first determination process for comparing the calculated level difference and a preset threshold value for each frequency component and determining a frequency component for which the level difference is greater than the threshold value to be a frequency component for which the target signal is dominant;

a process for extracting a frequency component for which it is determined that the target signal is dominant based on the first determination process from the first amplitude spectrum; and
a time waveform conversion process for converting the extracted first amplitude spectrum to a time waveform using the first phase spectrum.

(CLAIM 14) The recording medium according to Claim 13; wherein the program comprises a program for making the computer execute the following processes:

a noise suppression process for applying a preset weighting factor to the first amplitude spectrum of a frequency component for which it is not determined that the target signal is dominant in the first determination process; and
also converting a frequency component to which the weighting factor is applied to a time waveform using the first

phase spectrum of a corresponding frequency in the time waveform conversion process.

(CLAIM 15) The recording medium according to Claim 13 or 14; wherein the program comprises a program for making the computer execute the following processes:

a second determination process for comparing a magnitude of a first amplitude spectrum of a frequency component for which it is determined that the target signal is dominant and a preset silent section determination threshold value and determining that the sound source of the target signal is in a silent state when the first amplitude spectrum is smaller than the silent section determination threshold value; and
a process which, when it is determined by the second determination process that the sound source of the target signal is in a silent state, updates the threshold value used in the first determination process to at least the level difference.

(CLAIM 16) The recording medium according to Claim 14 or 15; wherein the program comprises a program for making the computer execute the following process:

a process for calculating a weighting factor to be applied in the noise suppression process by comparing a magnitude correlation of the level difference of each frequency component and the threshold value used in the first determination process.

(CLAIM 17) The recording medium according to Claim 15 or 16; wherein the program comprises a program for making the computer execute the following process:

a process for attenuating a frequency component which is extracted after it is determined that the target signal is dominant and the output of the noise suppression process or the output of the time waveform conversion process when it is determined by the second determination process that the sound source of the target signal is in a silent state.

(CLAIM 18) A voice recognition device comprising:

a first microphone installed at a position close to a sound source of a target signal (voice signal);

a second microphone installed at a position farther from the sound source of the target signal than said position;

first and second spectrum conversion means for converting respective output signals of the first and second microphones to first and second amplitude spectra and first and second phase spectra;

a level difference calculation means for calculating a level difference in the first and second amplitude spectra for each corresponding frequency component outputted from the first and second spectrum conversion means;

a sound source frequency component selection means for comparing the level difference for each frequency component outputted from the level difference calculation means and a preset threshold value, and determining a frequency component for which the level difference is greater than the threshold value to be a frequency component for which the target signal is dominant;

a sound source frequency component extraction means for extracting a frequency component for which it is determined that the target signal is dominant based on said determination from the first amplitude spectrum of an output signal of the first microphone;

a time waveform conversion means for converting the first

amplitude spectrum extracted from the sound source frequency component extraction means to a time waveform using the first phase spectrum;

a voiced state determination means for determining whether the sound source of the target signal is in a voiced state or a voiceless state;

a voice accumulation means for accumulating the output of the time waveform conversion means;

a voice recognition processing means for performing voice recognition using the output of the voice accumulation means; and

a voice recognition control means for performing control to input only voice data of voiced sections into the voice recognition processing means using the output results of the voiced state determination means and control to output the recognition results processed by the voice recognition processing means.

(CLAIM 19) The voice recognition device according to Claim 18, comprising:

a noise frequency component suppression means for applying a preset weighting factor to a frequency component for which it is not determined that the target signal is dominant in the sound source frequency component selection means;

wherein in the time waveform conversion means, the frequency components processed by the sound source frequency component extraction means and the noise frequency component suppression means are converted to time waveforms using the first phase spectrum.

(CLAIM 20) The voice recognition device according to Claim 18 or 19, comprising:

a threshold value calculation means which, when it is determined by the voiced state detection means that the sound source of the target signal is in a silent state, calculates and updates the threshold value used in the sound source frequency component selection means to at least the level difference outputted from the level different calculation means.

(CLAIM 21) The voice recognition device according to Claim 19 or 20, comprising:

a weighting factor calculation means for calculating a weighting factor to be applied to a frequency component for which it is not determined that the target signal is dominant in the noise frequency component suppression means by comparing a magnitude correlation of the level difference of each frequency component from the level different calculation means and the threshold value of the sound source frequency component selection means.

(CLAIM 22) The voice recognition device according to Claim 18, 19, 20, or 21;

wherein the first microphone and the second microphone are incorporated into a handset, a headset, an ear microphone set, or the like so that the latter is situated at a position closer to the mouth of a user than the latter during use.

(CLAIM 23) The voice recognition device according to any one of claims 18 to 22;

wherein the voiced state determination means comprises a voicing detection means for detecting a voiced state of a sound source of a target signal, and determines whether the sound source of the target signal is in a voiced state or a

voiceless state based on the output results of the voicing detection means.

(CLAIM 24) The voice recognition device according to any one of claims 18 to 22;

wherein the voiced state determination means is a means for comparing a magnitude of the output amplitude spectrum of the sound source frequency component extraction means and a preset silent section determination threshold value and determining that the sound source of the target signal is in a voiceless state when the amplitude spectrum is smaller than the silent section determination threshold value.

(CLAIM 25) A voice recognition method comprising:

installing a first microphone at a position close to a sound source of a target signal (voice signal);

installing a second microphone at a position farther from the sound source of the target signal than said position;

converting respective output signals of the first and second microphones to first and second amplitude spectra and first and second phase spectra;

calculating a level difference in the first and second amplitude spectra for each corresponding frequency component;

comparing the calculated level difference for each frequency component and a preset threshold value, and determining a frequency component for which the level difference is greater than the threshold value to be a frequency component for which the target signal is dominant with a first determination means;

extracting a frequency component for which it is determined by the first determination means that the target signal is dominant from the first amplitude spectrum;

converting the first amplitude spectrum of a frequency component for which the target signal is dominant to a time waveform with a time waveform conversion means using the first phase spectrum;

determining whether the sound source of the target signal is in a voiced state or a voiceless state;

accumulating the output of the time waveform conversion means in a voice accumulation means;

a voice recognition processing means for performing voice recognition with a voice recognition processing means using the output of the voice accumulation means; and inputting only voice data of voiced sections into the voice recognition processing means and outputting the recognition results processed by the voice recognition processing means.

(CLAIM 26) The voice recognition method according to Claim 25;

wherein a preset weighting factor is applied to the first amplitude spectrum of a frequency component for which it is not determined that the target signal is dominant in the comparison with the threshold value; and

in the time waveform conversion means, a frequency component to which the weighting factor is applied is also converted to a time waveform using the first phase spectrum of a corresponding frequency.

(CLAIM 27) The voice recognition method according to Claim 25 or 26;

wherein when it is determined that the sound source of the target signal is in a silent state, the threshold used in the first determination means is updated to at least the level difference.

(CLAIM 28) The voice recognition method according to Claim 26 or 27;

wherein the weighting factor to be applied to a frequency component for which it is not determined that the target signal is dominant is calculated by comparing a magnitude correlation of the level difference of each frequency component and the threshold value of the first determination means.

(CLAIM 29) The voice recognition method according to Claim 25, 26, 27, or 28;

wherein the first microphone and the second microphone are incorporated into a handset, a headset, an ear microphone set, or the like so that the former is situated at a position closer to the mouth of a user than the latter during use.

(CLAIM 30) The voice recognition method according to any one of Claims 25 to 29;

wherein the voiced state determination comprises a voicing detection means for detecting a voiced state of a sound source of a target signal, and determines whether the sound source of the target signal is in a voiced state or a voiceless state based on the output results of the voicing detection means.

(CLAIM 31) The voice recognition method according to any one of Claims 25 to 29;

wherein the voiced state determination compares a magnitude of a first amplitude spectrum for which it is determined that the target signal is dominant and which is extracted based on the determination of the first determination means, and a preset silent section determination threshold value, and determines that the sound source of the target signal is in a voiceless state when the first amplitude spectrum is smaller than the silent section determination threshold value.

(CLAIM 32) A recording medium for recording a program for making a computer of a sound collection device execute the following processes:

a process for converting respective output signals of a first microphone installed at a position close to a sound source of a target signal (voice signal) and a second microphone installed at a position farther from the sound source of the target signal than said position to first and second amplitude spectra and first and second phase spectra;

calculating a level difference in the first and second amplitude spectra for each frequency;

a first determination process for comparing the calculated level difference and a preset threshold value and determining a frequency component for which the level difference is greater than the threshold value to be a frequency component for which the target signal is dominant;

a process for extracting a frequency component for which it is determined that the target signal is dominant based on the first determination process from the first amplitude spectrum;

a time waveform conversion process for converting the first amplitude spectrum of the extracted frequency component to a time waveform using the first phase spectrum of a corresponding frequency;

a second determination process for determining whether the sound source of the target signal is in a voiced state or a voiceless state;

a process for accumulating the output of the time waveform conversion process in a voice accumulation means;

a process for performing voice recognition with a voice recognition processing means using the output of the voice accumulation means; and

a process for inputting only voice data in voiced sections into the voice recognition processing means using the results of the second determination process and outputting the recognition results processed by the voice recognition processing means.

(CLAIM 33) The recording medium according to Claim 32, wherein the program comprises a program for making the computer execute the following processes:

a process for applying a preset weighting factor to a frequency component for which it is not determined that the target signal is dominant in the first determination means; and

also converting a frequency component to which the weighting factor is applied to a time waveform using the first phase spectrum of a corresponding frequency in the time waveform conversion process.

(CLAIM 34) The recording medium according to Claim 32 or 33, wherein the program comprises a program for making the computer execute the following process:

a process which, when it is determined by the second determination process that the sound source of the target signal is in a silent state, updates the threshold value used in the first determination process to at least the level difference.

(CLAIM 35) The recording medium according to Claim 33 or 34, wherein the program comprises a program for making the computer execute the following process:

a process for calculating a weighting factor to be applied to a frequency component for which it is not determined that the target signal is dominant by comparing a magnitude correlation of the level difference and the threshold of the first determination process.

(CLAIM 36) The recording medium according to any one of Claims 32 to 35;

wherein the second determination process determines, based on the output results of a voicing detection means which detects a voiced state of a sound source of a target signal, whether the sound source of the target signal is in a voiced state or a voiceless state.

(CLAIM 37) The recording medium according to any one of Claims 32 to 35;

wherein the second determination process compares a magnitude of a first amplitude spectrum for which it is determined that the target signal is dominant and which is extracted based on the determination of the first determination process, and a preset silent section determination threshold value, and determines that the sound source of the target signal is in a voiceless state when the first amplitude spectrum is smaller than the silent section determination threshold value.

(DETAILED DESCRIPTION OF THE INVENTION)

(0001)

(TECHNICAL FIELD OF THE INVENTION) The present invention relates to a sound collection device for suppressing ambient noise components in a sound source signal containing ambient noise and extracting a target signal, a voice recognition device which has the sound collection function of the sound collection device and achieves a high recognition rate, methods thereof, and a program recording medium.

(0002)

(CONVENTIONAL TECHNOLOGY) One conventional technique for extracting a target signal with a good SN ratio in noise is a spectral subtraction, wherein the amplitude spectrum of a noise signal (estimated value) is subtracted from the amplitude spectrum of a target signal over which noise is superimposed (for example, "S.F. Boll, 'Suppression of acoustic noise in speech using spectral subtraction', IEEE Trans., ASSP, Vol. 27, No. 2, p113-120, (1979)"). FIG. 20 is a flowchart for explaining an example of a conventional technique. First, an output signal L of an air conduction microphone (which receives acoustic signals using air as a medium) is converted to a spectrum (S01). Next, a spectrum $S(\omega_i)$ from which noise is removed is found by subtracting the amplitude spectrum of estimated noise from the amplitude spectrum of the output signal L of the air conduction microphone (S02). The spectrum $S(\omega_i)$ is then converted to a time waveform using the phase spectrum of the output signal L of the air conduction microphone to obtain a target signal in which noise is suppressed (S03).

(0003) Examples of methods used to estimate a noise spectrum in the process described above include ① a method of finding a noise spectrum by installing a microphone for noise collection, and ② a method of using a noise spectrum obtained from a silent section of the target sound source. However, these noise estimation methods include the following such problems, which induce noise spectrum estimation error.

(0004)

(PROBLEM TO BE SOLVED BY THE INVENTION) In the method of ①, when the positions at which the target signal collection microphone and the noise collection microphone are disposed are too far apart, differences arise in the noise spectrum characteristics picked up by the two microphones. Conversely, when the positions where the two microphones are configured are too close, the contribution of the target signal is included in the noise collection microphone. That is, the method of ① has the problem that it is difficult to establish a configuration for the two microphones for estimating an accurate noise spectrum.

(0005) In the method of ②, it is impossible to avoid estimation error in the noise spectrum arising when the spectral characteristics of noise change in voiced sections of the target sound source. In particular, in intense non-stationary noise such as traffic noise, the noise suppression process cannot be expected to yield good results. It is thus presently difficult to accurately estimate a noise spectrum with current spectral subtraction methods.

(0006) When performing voice recognition in noise, an effective means of for dealing with variation in voice signals due to ambient noise is to perform noise removal processing at the voice input stage (for example, "Acoustics of ambient noise in voice recognition", Journal of the Acoustical Society of Japan, Vol. 51, No. 4, p331-335, (1995))

However, a technique that can be applied in any environment as described above has not yet been established for current noise suppression processing.

(0007) In addition, in order to perform voice recognition, it is necessary to first detect voiced sections. However, it is difficult to accurately detect voiced sections in noise. Although a method of performing section detection from a signal from which noise has been removed with the spectral subtraction method or the like described above may be

considered, implementing this method requires the realization of accurate noise spectrum estimation, which is a current problem.

(0008) An object of the present invention is to provide a novel sound collection device having a noise suppression processing function and a novel voice recognition device having the noise suppression processing function and an accurate voice section detection function in noise, which do not require the accurate estimation of a noise signal and can be applied even to non-stationary noise.

(0009)

(MEANS FOR SOLVING THE PROBLEM) A first mode of the present invention is provided with a first microphone installed at a position close to a sound source of a target signal and a second microphone installed at a position farther from the sound source of the target signal than the first position, wherein respective output signals of the first and second microphones are converted to amplitude spectra and phase spectra by first and second spectrum conversion means, and a level difference is calculated by a level difference calculation means for the amplitude spectrum of each frequency component outputted from the first and second spectrum conversion means. This level difference for each frequency component is compared to a preset threshold value by a sound source frequency component selection means, and a frequency component for which the level difference is greater than the threshold value is determined by the sound source frequency component selection means to be a frequency component for which the target signal is dominant. A frequency component for which it is assessed that the target signal is dominant is extracted by a sound source frequency component extraction means from the output amplitude spectrum of the first spectrum conversion means. The amplitude spectrum of this extracted frequency component for which the target signal is dominant is converted by a time waveform conversion means to a time waveform using the phase spectrum calculated by the first spectrum conversion means.

(0010) A second mode of the present invention is one in which, in the sound collection device of the first mode, a preset weighting factor (smaller than 1) is applied by a noise frequency component suppression means to an amplitude spectrum of the first spectrum conversion means associated with a frequency component for which it is not determined by the sound source frequency component selection means that the target signal is dominant, and the product output is converted by the time waveform conversion means to a time waveform using the phase spectrum of the corresponding frequency component calculated by the first spectrum conversion means.

(0011) A third mode of the present invention is one in which, in the sound collection device of the first or second mode, a magnitude of an amplitude spectrum of a frequency component for which it is determined by the sound source frequency component extraction means that the target signal is dominant and a preset silent section determination threshold value are compared by a sound source silent section determination means, and it is determined that the sound source of the target signal is in a silent state when the amplitude spectrum is smaller than the silent section determination threshold value. When it is determined that the sound source of the target signal is in a silent state, the threshold value used in the sound source frequency

component selection means is calculated and updated by a threshold value calculation means to at least the level difference outputted by the level difference calculation means. (0012) A fourth mode of the present invention is one in which, in the sound collection device of the second or third mode, the weighting factor to be applied by the noise frequency component suppression means to a frequency component for which it is not determined that the target signal is dominant is calculated by a weighting factor calculation means by comparing a magnitude correlation of the level difference of each frequency component determined by the level difference calculation means and the threshold value of the sound source frequency component selection means.

(0013) A fifth mode of the present invention is one in which, in the sound collection device of the third or fourth mode, when it is determined by the sound source silent section determination means that the sound source of the target signal is in a silent state, the outputs of the sound source frequency component extraction means and the noise frequency component suppression means or the output of the time waveform conversion means is attenuated by a sound source silent section attenuation means. A sixth mode of the present invention is one in which, based on the premise of any one of the first through fourth modes, a voiced state of a sound source of a target signal is detected by a voicing detection means, and it is determined by a voiced state determination means whether the sound source of the target signal is in a voiced state or a voiceless state from the detection results. This voiced state determination means may use the sound source silent section determination means or may be provided independently. Signal converted to time waveforms are accumulated in a voice accumulation means, and voice recognition is performed by a voice recognition processing means using the accumulated signals. Only voice data in voiced sections are inputted into the voice recognition processing means by a voice recognition control means using the determination results of the voiced state determination means, and the recognition results are outputted after recognition processing.

(0014) A seventh mode of the present invention is one in which, in the any one of the first through sixth modes, the first microphone installed at a position close to the sound source of the target signal and the second microphone installed at a position farther from the sound source of the target signal are incorporated into a handset, a headset, an ear microphone set, or the like so that the former is situated at a position closer to the mouth of a user than the latter during use.

(0015) "EFFECT" In the first mode, it is determined whether a frequency component is one in which the target signal is dominant based on a level difference in the amplitude spectra of microphones installed at a position close to the sound source of the target signal and a position farther from the sound source of the target signal than the first position. In the device configuration of this invention, the positional relationship between the sound source and each microphone varies very little for a target signal, and the distances are also short, so a stable level difference arises between the two microphones. On the other hand, for a noise signal, the distance between the sound source and the microphones may be considered longer than the distance between the sound source of the target signal and the microphones. Therefore, the output level difference between the two microphones arising due to the target signal is always larger than the level difference of the output level difference of the two microphones arising due to noise. In this invention, focusing

attention on the fact that the level difference arising in the two microphones differs for the target signal and noise, as described above, a process for extracting a frequency component for which the target signal is dominant is performed. Such a frequency component extraction process enables a noise suppression process which does not require the accurate estimation of a noise spectrum.

(0016) In the second mode, a corresponding frequency component is attenuated by applying a preset weighting factor to a frequency component for which it is assessed that the target signal is dominant in the first mode. With this noise frequency component suppression process, it is possible to prevent sound quality degradation of the processed signal due to the dropping of the target signal component superimposed with a frequency component for which it is not determined that the target signal is dominant, which makes it possible to improve sound quality.

(0017) In the third mode, when it is determined that the target sound source is in a silent state in the first and second mode, the set threshold value is calculated, and the precision with which a frequency component for which the target signal is dominant is determined by the sound source frequency component selection means is improved, and the sound quality improves. In the fourth mode, the amount of attenuation in a frequency component for which the target signal is not dominant in the second or third mode is controlled in accordance with the degree of noise superimposition, and the sound quality improves.

(0018) In the fifth mode, when it is determined that the sound source of the target signal is in a silent state in the third or fourth mode, the outputs of the sound source frequency component extraction means and the noise frequency component suppression means or the output of the time waveform conversion means is attenuated, and the noise suppressing effect improves. In the sixth mode, focusing attention on the fact that the level difference arising in the two microphones differs for the target signal and noise, a process is performed to extract a component for which the target signal is dominant, as in the first mode, which enables noise suppression processing which does not require the accurate estimation of a noise spectrum, and a voice recognition process using a voice signal with an enhanced SN ratio becomes possible as a result.

(0019) In addition, for the detection of voiced sections, the output of a bone conduction microphone, the output of an external auditory meatus microphone, or the results of analysis of a mouth image captured by a camera, for example, are used. A bone conduction sound or an external auditory meatus sound is not of sufficient sound quality to attenuate high-frequency components, but the noise resistance is excellent, so there is minimal transformation in the output signals due to the effects of noise. FIG. 21 is a drawing comparing the time waveforms of air conduction sound and bone conduction sound when a voice is generated in a laboratory environment in noise (pink noise) of 90 dBA. It can be confirmed from this drawing that a bone conduction microphone yields clearer voiced/voiceless sections than an air conduction microphone. Using an output signal of this bone conduction microphone or external auditory meatus microphone enables the detection of voiced sections with high precision in noise. In addition, in a system using the video of a camera such as a television phone, using video information of the mouth of the speaker enables the detection of voiced sections independent of ambient noise.

(0020) Under the premise of the second mode, the noise frequency suppression process prevents sound quality degradation of the processed signal due to the dropping of the target signal component superimposed with a frequency component for which it is not determined that the target signal is dominant, which makes it possible to improve sound quality. Under the premise of the third mode, the voiceless section threshold value calculation means enhances the precision with which a frequency component for which the target signal is dominant is determined by the sound source frequency component selection means is improved, and also prevents sound quality degradation, which improves voice recognition precision.

(0021) Under the premise of the fourth mode, the weighting factor calculation means calculates the amount of attenuation in a frequency component for which the target signal is not dominant in accordance with the degree of noise superimposition, and also prevents sound quality degradation, which improves voice recognition precision. In the seventh mode, the first microphone installed at a position close to the sound source of the target signal and the second microphone installed at a position farther from the sound source of the target signal are incorporated into a handset, a headset, an ear microphone set, or the like so that the former is situated at a position closer to the mouth of a user than the latter during use, which enhances the noise resistance performance of a transmission signal in each handset and makes it possible to improve voice recognition precision when using voice commands or the like.

(0022)

(EMBODIMENTS OF THE INVENTION) "Example 1" FIG. 1 is a block diagram illustrating an example of the first mode of the present invention. A microphone 1 is for a nearby sound source and is disposed at a position closer to a sound source of a target signal than a microphone 2 (for a remote sound source). The level difference in the outputs of the microphones 1 and 2 for a sound from the sound source of the target signal is large, but a sound source of noise is sufficiently far from the microphones 1 and 2, and the level difference in the outputs of the microphones 1 and 2 due to noise is zero or minimal.

(0023) FIG. 7 is a flowchart illustrating the processing flow of the example of the first mode of the present invention. First, a target signal over which noise is superimposed is captured by each of the microphones 1 and 2, and these are read as a digital signal (S02). The signals read by the microphones 1 and 2 are called L and R hereinafter. In spectrum conversion parts 3 and 4, the captured signals L and R are converted to spectra $L(\omega_i)$ and $R(\omega_i)$ (S03). This conversion is executed by a discrete Fourier transform, for example.

(0024) In a level difference calculation part 5, the level difference $\Delta LR(\omega_i)$ given by the following formula is calculated for each of the frequency components of $L(\omega_i)$ and $R(\omega_i)$ (S04).

$$\Delta LR(\omega_i) = 20 \log_{10}(|L(\omega_i)|/|R(\omega_i)|)$$

In the formula, ω_i is the frequency ($i=1, 2, \dots, n$), and $|L(\omega_i)|$ and $|R(\omega_i)|$ are respectively the amplitude spectrum components of the L and R signals.

(0025) In a voice frequency component selection part 6, a frequency for which the target signal is dominant is selected based on the magnitude correlation of $\Delta LR(\omega_i)$ and a preset

threshold value $Th(\omega_i)$ for each frequency component. The conditions for determining whether the frequency is one for which the target signal is dominant are determined by the following formulas (S05).

$\Delta LR(\omega_i) > Th(\omega_i) \rightarrow$ the target signal is dominant

$\Delta LR(\omega_i) \leq Th(\omega_i) \rightarrow$ the target signal is not dominant

In a sound source frequency component extraction part 7, only a frequency component for which the target signal is dominant is extracted with respect to a frequency selected by the sound source frequency component selection part 6 (S06).

(0026) Finally, the output of (S06) is converted to a time transformation using a phase spectrum $\Phi(\omega_i)$ of the signal L in a time waveform conversion part 8, and a time waveform signal is outputted (S07). The threshold value $Th(\omega_i)$ can be determined experimentally from the correlation between the microphones 1 and 2 and the target signal sound source (for example, the mouth of the user of the microphones 1 and 2).

"Example 2" In the first mode, frequency components for which it is not determined that the target signal is dominant are completely removed. However, when components formed by the mutual overlapping of voice and noise frequency components are also completely removed, this may cause distortion in sound quality, which is not preferable. The second mode improves the sound quality after processing by overcoming the aforementioned problem of "excessive extraction of noise components" for frequency components combining a voice and noise.

(0027) FIG. 2 is a block diagram illustrating an example of the second mode. In FIG. 2, the operation is the same as in FIG. 1 with the exception that a noise frequency component suppression part 9 is connected to the output side of each of the spectrum conversion part 3 and the sound source frequency component selection part 6. The processing of the noise frequency component suppression part 9 will be described hereinafter using the flowchart of FIG. 8 illustrating an example of the second mode. In the noise frequency component suppression part 9, a value obtained by applying a weighting factor $w(\omega_i)$ set externally to a frequency component for which it is determined by the sound source frequency component selection part 6 that the target signal is not dominant – that is, the frequency component $L(\omega_i)$ being outputted by the spectrum conversion part 3, which satisfies the relation $\Delta LR(\omega_i) \leq Th(\omega_i)$ – is defined as the spectrum component $S(\omega_i)$ after the noise suppression process (S07).

(0028) $S(\omega_i) = w(\omega_i) \times L(\omega_i)$

The value of $w(\omega_i)$ should be set to a small value ($w(\omega_i) \neq 0$) in a range in which the amount of noise suppression is to be emphasized, and should be set to a large value ($w(\omega_i) \neq 1$) in a range in which sound quality is to be emphasized. A frequency component for which it is determined that the target signal is dominant is processed by the sound source frequency component extraction part 7 in the same manner as in the example of the first mode, and the extracted frequency component is defined as $S(\omega_i)$ (S06).

(0029) $S(\omega_i) = L(\omega_i)$

The output $S(\omega_i)$ of the noise frequency component suppression part 9 and the sound source frequency component extraction part 7 is subjected to an inverse Fourier transform in the time waveform conversion part 8, and a time waveform signal is outputted (S08).

"Example 3" In the first and second modes, the threshold

value $Th(\omega_i)$ used in the sound source frequency component selection part 6 to determine whether the target signal is dominant in a given frequency component is set externally. The third mode calculates and corrects the threshold $Th(\omega_i)$ using the level difference in each frequency component between the microphones 1 and 2 arising in ambient noise when the target sound source is in a silent state, thereby enhancing the precision with which the sound source frequency component selection part 6 determines whether the target signal is dominant and improving sound quality.

(0030) FIG. 3 is a block diagram illustrating an example of the third mode. In FIG. 3, the operation is the same as in FIG. 2 with the exception that a sound source silent section determination part 10 is connected to the output side of the sound source frequency component extraction part 7 and that a sound source silent section threshold value calculation part 11 is connected to the output side of the sound source silent section determination part 10. The processing of the sound source silent section determination part 10 and the sound source silent section threshold value calculation part 11 will be described hereinafter using the flowchart of FIG. 9 illustrating an example of the third mode.

(0031) In the sound source silent section determination part 10, a sum P of amplitude spectra for which the target signal is dominant is first found from the sound source frequency component extraction part 7 (S08), and the silent state of the sound source of the target signal is detected from the magnitude correlation of P and a threshold value PTh which is set externally (S09).

$P > PTh \rightarrow$ the sound source of the target signal is in a voiced state

$P \leq PTh \rightarrow$ the sound source of the target signal is in a voiceless state

When it is determined that the sound source of the target signal is in a voiceless state in the sound source silent section determination part 10, the threshold value $Th(\omega_i)$ is calculated by the sound source silent section threshold value calculation part 11. For example, a new threshold value is calculated by the following formula (S10, S11). $Th(\omega_i) = \Delta LR(\omega_i)$ (only when $\Delta LR(\omega_i) > Th(\omega_i)$). That is, the frequency components of the outputs of the microphones 1 and 2 are ordinarily roughly equal with respect to noise, and $\Delta LR(\omega_i)$ is zero, but $\Delta LR(\omega_i)$ is relatively large for frequency components due to reflection or the like. For the $\Delta LR(\omega_i)$ in this sound source silent section, even when the target signal is present, a noise frequency component is present, and the threshold value is increased by this component so that noise is effectively removed. Note that $Th(\omega_i)$ may be reset as necessary when it surpasses a certain magnitude so that the $\Delta LR(\omega_i)$ of noise that has increased temporarily does not have a permanent effect.

(0032) "Example 4" In the second and third modes, the weighting factor $w(\omega_i)$ to be applied to a frequency component for which the target signal is not dominant in the noise frequency component suppression part 9 is set externally. The fourth mode calculates the weighting factor $w(\omega_i)$ using the level difference $\Delta LR(\omega_i)$ of each frequency component calculated by the level difference calculation part 5 and the threshold $Th(\omega_i)$ so as to control the amount of attenuation of each frequency component in the noise frequency component suppression part 9 in accordance with the amount of superimposition of noise components, thereby improving sound quality.

(0033) FIG. 4 is a block diagram illustrating an example of the fourth mode. The operation is the same as in FIG. 3 with

the exception that a weighting factor calculation part 12 is connected to the output sides of each of the level difference calculation part 5 and the sound source frequency component selection part 6. The processing in the weighting factor calculation part 12 will be described hereinafter using the flowchart of FIG. 10 illustrating an example of the fourth mode. In the weighting factor calculation part 12, the weighting factor $w(\omega_i)$ to be applied to a frequency component for which it is determined that the target signal is not dominant in the sound source frequency component selection part 6 is calculated by the following formula, for example (S07).

(0034)

$$w(\omega_i) = (\Delta LR(\omega_i) / Th(\omega_i))^{npw}$$

The value of npw may be set to a large value when it is desirable to increase the noise suppressing effect. The $w(\omega_i)$ found from this formula assumes a small value when the degree of superimposition of noise is large because $\Delta LR(\omega_i)$ is governed by the level difference of the noise spectra, and it assumes a large value when the degree of superimposition of noise is small because $\Delta LR(\omega_i)$ is governed by the level difference in the target signal spectra. Note that when the value of $w(\omega_i)$ calculated by the formula above is negative, it is considered to be a frequency component for which noise is dominant, and the value is set to a small value such as 0, for example.

(0035) "Example 5" The fifth mode is one in which, when it is determined by the sound source silent section determination part 10 that the sound source of the target signal is in a silent state in the third or fourth mode, the outputs of the sound source frequency component extraction part and the noise frequency component extraction part or the output of the time waveform conversion part are attenuated so as to enhance the noise suppressing effect.

(0036) FIG. 5 is a block diagram illustrating an example of the fifth mode. In FIG. 5, the operation is the same as in FIG. 4 with the exception that a sound source silent section attenuation part 13 is inserted between the noise frequency component suppression part 9 and the time waveform conversion part 8. The processing in the sound source silent section attenuation part 13 will be described hereinafter using the flowchart of FIG. 11 illustrating an example of the invention of the fifth mode.

(0037) In the sound source silent section attenuation part 13, when it is assessed by the sound source silent section determination part 10 that the sound source of the target signal is in a silent state (S10), the output $S(\omega_i)$ of the sound source frequency component extraction part 7 and the noise frequency component suppression part 9 is attenuated over the entire spectrum (S13). Note that the processing of the sound source silent section attenuation part 13 may also be performed for the output signal of the time waveform conversion part 8, and the effect thereof is the same as the effect of the processing of (S13).

(0038) "Experimental Example 1" An experimental example to which the fourth mode is applied will be described hereinafter. Using ambient noise in which the target signal is a voice and the noise signal is a station platform, signals of the microphones 1 and 2 were created with a computer as illustrated in FIG. 12. That is, the target signal (voice) is inputted into the microphone 2 after being reduced by 6 dB in comparison to the microphone 1, and noise is inputted at the same level for the microphones 1 and 2. The S/N ratio is defined as the ratio of the average power of the target signal to the average power of the noise signal, and the value was set

to -9 dB in the microphone 1. the frequency resolution in spectral decomposition of the signal was 22 Hz, the analysis frames were 46 ms, and the updating period of the analysis frames was 23 ms.

(0039) FIG. 13 illustrates a target signal A prior to processing by the microphone 1, a noise signal B, a noise signal + target signal C, and a signal D after processing. It can be confirmed from this drawing that the signal after process faithfully restores the target signal before processing.

"Example 6" FIG. 14 illustrates an example of the sixth embodiment, wherein portions corresponding to those of FIG. 1 are labeled with the same symbols. In this example, the output of a voicing detection part 14 is supplied to a voiced state determination part 15, and the output of the voiced state determination part 15 is inputted into a voice recognition control part 16. The output of the time waveform conversion part 8 is accumulated in a voice accumulation part 17, and the accumulated signals are subjected to voice recognition processing by a voice recognition processing part 18. The recognition results are outputted via the voice recognition control part 16.

(0040) FIG. 15 illustrates the processing procedure of an example of the sixth mode. In this process, steps S01 to S06 are the same as in the process illustrated in FIG. 7. In step S05, a frequency component for which a target signal of $\Delta LR(\omega_i) \leq Th(\omega_i)$ is not dominant is considered a frequency component for which noise is dominant, and the value is changed to a small value eps (S07). The fact that the value of eps is 0, for example, and is converted to a time waveform in step S08 is the same as in FIG. 7.

(0041) In the voicing detection part 14, a bone connection or external auditory meatus microphone or a camera can be connected and operated. FIG. 15 illustrates a method of determining the voiced/voiceless state of a sound source of a target signal with the voiced state determination part 15. First, the power PB of an output signal of the bone connection or external auditory meatus microphone is found (S09). Next, the voiced/voiceless state of the sound source of the target signal is determined from the magnitude correlation of PB and a value PTHB which is set externally (S10).

(0042)

PB > PTHB → the sound source of the target signal is in a voiced state

PB ≤ PTHB → the sound source of the target signal is in a voiceless state

Note that when a camera is used for the voicing detection part 14, the voiced state determination may be performed by monitoring fluctuation in a video signal of the mouth, for example. The voice recognition control part 16 controls two elements: ① voice data input from the voice accumulation part 17 into the voice recognition processing part 18, and ② output of voice recognition results. The control of ① may be achieved by two methods indicated by ①-1 and ①-2.

(0043) ①-1: Method of accumulating only the data of the time waveform conversion part 8 corresponding to voiced sections in the voice accumulation part 17 and inputting all of the data of the voice accumulation part 17 into the voice recognition processing part 18

①-2: Accumulating all of the data of the time waveform conversion part 8 in the voice accumulation part 17 and inputting the data corresponding to voiced sections into the voice recognition processing part 18 from among the data

In this example, a specific example of the control method of ①-1 will be described.

(0044) In the voice recognition control part 16, a data input terminal of the voice accumulation part 17 is turned ON at the moment when voicing starts – that is, the moment when the voiced state determination part 15 determines that a preprocessing frame is in a voiceless state and that the current frame is in a voiced state (S10, S11). The accumulation of the output signals of the time waveform conversion part 8 in the voice accumulation part 17 starts from this moment (S12). This accumulation is performed while it is determined by the voiced state determination part 15 that the frame is in a voiced state, and it ends at the moment when the frame transitions from a voiced state to a voiceless state (the voiced state determination part 15 determines that the preprocessing frame is in a voiced state and that the current frame is in a voiceless state (S13)) (S14).

(0045) When the accumulation of voice data by the voice accumulation part 17 is complete (S14), the data accumulated by the voice accumulation part 17 is inputted into the voice recognition processing part 18, which receives a command from the voice recognition control part 16 and performs voice recognition processing (S15). The voice recognition results are outputted to an external device via the voice recognition control part 16 (equivalent to the control of ② described above). The voice data accumulated in the voice accumulation part 17 is cleared immediately after being outputted to the voice recognition processing part 18 (S16).

(0046) Note that when the voice recognition algorithm that is employed uses only voice spectrum information and a voice time waveform is unnecessary, the processed voice data accumulated in a storage device or the like may be the processed spectra (S06) and (S07) rather than the time waveform output of (S08). In this case, the time waveform conversion process of (S08) is unnecessary.

"Example 7" The configurations 14 to 18 of Example 6 are added to Example 2 to overcome the aforementioned problem of "excessive extraction of noise components" for frequency components combining a voice and noise., which enhances sound quality after processing and improves the precision of voice recognition. FIG. 2 illustrates a case in which the configurations 14 to 18 of FIG. 14 are added, and steps S09 to S18 of FIG. 15 are added after S08 in FIG. 8.

(0047) "Example 8" The configurations 14 to 18 of Example 6 are added to Example 3, wherein the threshold value $Th(\omega_i)$ is calculated and corrected using the level difference in each frequency component between the microphones 1 and 2 arising in ambient noise when the sound source of the target signal is in a silent state (voiceless section), which improves the precision when determining whether the target signal is dominant in the sound source frequency component selection part 6 and prevents sound quality degradation so as to improve voice recognition precision. In this case, as illustrated in FIG. 16, the configurations 14 to 18 of FIG. 14 are added to FIG. 3, and the sound source silent section determination part 10 in FIG. 3 is omitted so that the output of the voiced state determination part 15 is inputted into the voiceless (sound source silent) section threshold value calculation part 11. In the flow of the processing in this example, as illustrated in FIG. 17, when the voiced state determination part 15 determines that the sound source of the target signal is in a silent state with respect to FIG. 15 (S20), the threshold value $Th(\omega_i)$ is calculated by the voiceless section threshold value calculation part 11 (S17, S18). For example, a new value is calculated by the following formula (S17, S18), and the flow returns to step S02. $Th(\omega_i) = \Delta LR(\omega_i)$

(only when $\Delta LR(\omega_i) > Th(\omega_i)$). The voicing detection part 14 and the voiced state determination part 15 may be omitted, and sound source silent section determination part 10 may be used instead.

(0048) "Example 9" In this example, the weighting factor $w(\omega_i)$ is calculated using the level difference $\Delta LR(\omega_i)$ of each frequency component calculated by the level difference calculation part 5 and the threshold $Th(\omega_i)$ so as to control the amount of attenuation of each frequency component in the noise frequency component suppression part 9 in accordance with the degree of superimposition of noise components and to prevent sound quality degradation, thereby improving voice recognition precision. The weighting factor calculation part 12 illustrated in FIG. 4 is added to the configuration of FIG. 16, and step S07 in FIG. 10 is added to FIG. 17 accordingly in the flow of processing.

(0049) "Experimental Example 2" An experimental example to which Example 6 is applied will be described hereinafter. The output of a bone conduction microphone was used as an input signal of the voiced section detection part 14. The sound collection conditions are as illustrated in FIG. 18A – that is, the microphones are 150 cm away from the noise source, with a spacing of 10 cm between the first and second microphones – and the experimental conditions are illustrated in FIG. 18B.

(0050) FIG. 19 illustrates the results of finding the voice recognition rate of the output of the microphone 1 (before processing) and the output after the noise suppression process of Example 6 was performed (after processing). It can be seen from this drawing that the voice recognition rate in high noise is improved with a signal after processing in comparison to a signal before processing. FIG. 6 illustrates an example of the seventh mode. FIG. 6A is a case in which the microphones 1 and 2 are mounted to a handset 21. When the handset 21 is in use, the microphone 1 is positioned near the mouth 22 of the user – that is, near the sound source of the target signal – and the microphone 2 is positioned near the receiver portion of the handset – that is, near the ear 23.

(0051) FIG. 6B illustrates a case in which the microphones 1 and 2 are mounted to a headset 25, wherein during use while the headset 25 is affixed to the head 26 of the user, the microphone 2 is mounted to a portion of a receiver 27 in contact with the ear 23, and a support arm 28 is extended from this portion of the receiver 27. The free end of the support arm 28 is positioned near the mouth 22, and the microphone 1 is mounted thereon.

(0052) FIG. 6C illustrates a case in which the microphones are mounted to an ear microphone set 31, wherein in a state in which the ear microphone set 31 is mounted to a portion of the ear 23, an arm 32 is extended to the mouth 22 side, and the microphone 1 is mounted thereon. An arm 33 is extended to the opposite side as this arm 32, and the microphone 2 is mounted thereon.

(0053)

(EFFECT OF THE INVENTION) As described above, according to the invention of the first mode, a noise suppression process can be performed without estimating an accurate noise spectrum component by converting the respective output signals of a microphone installed at a position close to a sound source of a target signal and a microphone installed at a position farther from the sound source of the target signal than the first position to amplitude spectra and phase spectra, calculating a level difference in the amplitude spectra for each frequency component, comparing the level difference for each frequency component with a preset threshold value, determining a frequency component for which the level

difference is greater than the threshold value to be a frequency component for which the target signal is dominant, extracting these frequency components from the amplitude spectrum of the output signal of the microphone installed at a position near the sound source of the target signal, and converting the amplitude spectrum of the extracted frequency components for which the target signal is dominant to a time waveform using the phase spectrum calculated at the time of the spectrum conversion of the microphone near the sound source of the target signal. Accordingly, it is possible to obtain a sound collection device capable of extracting a target signal even when it is difficult to estimate the noise spectrum, as in the case of non-stationary noise or the like.

(0054) According to the invention of the second mode, it is possible to obtain a sound collection device capable of overcoming sound quality degradation due to the dropping of frequency components of the target signal that were present in a frequency band not extracted as a target signal, by applying a preset weighting factor to a frequency component for which it is not determined that the target signal is dominant in the sound collection device of the first mode to suppress noise frequency components, and also converting this suppressed component to a time waveform in the time waveform conversion using the phase spectrum calculated by the microphone near the sound source of the target signal.

(0055) According to the invention of the third mode, the magnitude of the amplitude spectrum of a frequency component for which it is determined by sound source frequency component extraction that the target signal is dominant is compared to a preset silent section determination threshold value in the sound collection device of the first or second mode, and it is determined that the sound source is in a silent state when the amplitude spectrum is smaller than the silent section threshold value. When it is determined that the sound source is in a silent state, the threshold value to be used in the sound source frequency component selection is updated to at least the level difference, so the precision of extracting a component for which the target signal is dominant improves in the sound source frequency component selection, and a sound collection device capable of enhancing the quality of signals after processing can be obtained.

(0056) In the invention of the fourth mode, the weighting factor to be applied to a frequency component for which it is not determined that the target signal is dominant during noise frequency component suppression is calculated by comparing the magnitude correlation of the level difference in each frequency component and the threshold at the time of the sound source frequency component selection in the sound collection device of the second or third mode, so it is possible to obtain a sound collection device capable of determining the amount of attenuation of each frequency component in accordance with the degree of superimposition of noise during noise frequency component suppression.

(0057) In the invention of the fifth mode, when it is determined that the sound source of the target signal is in a silent state in the sound collection device of the third or fourth mode, the sound source frequency component extraction output and the noise frequency component suppression output or the output of the time waveform conversion means is attenuated, so signals are attenuated when the sound source of the target signal is in a silent state, which suppresses noise and makes it possible to obtain a sound collection device with even less noise.

(0058) The invention of the sixth mode includes the roughly the same configuration as any one of the first through fourth

modes and further has a voice accumulation means and a voice recognition processing means, wherein the output results of the voiced state determination means are used to input only voice data in voiced sections and to output the recognition results processed by the voice recognition processing means, so a sound suppression process can be performed without estimating an accurate sound spectrum component. Accordingly, it is possible to obtain a voice recognition device capable of extracting a target signal even when it is difficult to estimate the noise spectrum, as in the case of non-stationary noise or the like, which enhances voice recognition precision. In addition, by using a bone conduction or external auditory meatus microphone or a camera for the detection of voiced sections, it is possible to detect voiced sections with minimal or no effects of ambient noise, which yields a voice recognition device with even better voice recognition precision.

(0059) In the invention of the seventh mode, the microphone installed at a position near the sound source of the target signal and the microphone installed at a position farther from the sound source of the target signal in the sound collection device or the voice recognition device of the first through sixth modes are incorporated into a handset, a headset, or an ear microphone set so that the former microphone is situated at a position closer to the mouth than the latter microphone during use. Therefore, the noise resistance performance of transmission signals can be improved in a conventional handset, headset, or ear microphone set, and the voice recognition precision can be improved when using voice commands or the like in a handset. Conventionally, a bone conduction microphone and a receiver are formed integrally as an ear microphone set for obtaining a transmission signal with excellent sound resistance. However, the frequency components of a voice captured by a bone conduction microphone are biased toward low-frequency components, and there are few high-frequency components, so the sound quality is poor. There is also a problem with acoustic coupling between the bone conduction microphone and the receiver. However, in the invention of the seventh mode, sound is collected based on air conduction, and the distance between the receiver and the microphone can be secured, so an ear microphone set free of the problems described above can be provided.

(0060) Note that the microphones used in the above explanation are not limited to omnidirectional microphones. For example, a microphone having directivity in the direction of the sound source of the target signal may be used for the microphone 1, and a microphone having directivity in the opposite direction as the sound source of the target signal may be used for the microphone 2. In this case, a super-directive microphone having sharp directivity only in the direction of the sound source of the target signal can be used.

(0061) The sound collection device of the present invention can be used not only for various sound collection devices which require sound suppression, but also for telephone devices intended for making calls or input devices for voice

recognition. In addition, FIGS. 1 to 5, 14, and 16 illustrate functional configurations, and each function can be achieved by a means primarily consisting of a microprocessor. That is, a function of reading out and decoding a program with a computer can also be achieved.

(BRIEF DESCRIPTION OF THE DRAWINGS)

FIG. 1 is a block diagram illustrating the functional configuration of an example of the invention of Claim 1.

FIG. 2 is a block diagram illustrating the functional configuration of an example of the invention of Claim 2.

FIG. 3 is a block diagram illustrating the functional configuration of an example of the invention of Claim 3.

FIG. 4 is a block diagram illustrating the functional configuration of an example of the invention of Claim 4.

FIG. 5 is a block diagram illustrating the functional configuration of an example of the invention of Claim 5.

FIG. 6 is a drawing illustrating an example of the schematic configuration of the seventh mode of the present invention.

FIG. 7 is a flowchart illustrating the operating procedure of an example of the invention of Claim 1.

FIG. 8 is a flowchart illustrating the operating procedure of an example of the invention of Claim 2.

FIG. 9 is a flowchart illustrating the operating procedure of an example of the invention of Claim 3.

FIG. 10 is a flowchart illustrating the operating procedure of an example of the invention of Claim 4.

FIG. 11 is a flowchart illustrating the operating procedure of an example of the invention of Claim 5.

FIG. 12 is a drawing illustrating an example of the application of signals and noise in a simulation experiment to which the invention of Claim 4 is applied.

FIG. 13 is a drawing illustrating a target signal prior to processing, a noise signal, a noise signal+target signal, and a signal after processing in an experimental example using the invention of Claim 4.

FIG. 14 is a block diagram illustrating the functional configuration of an example of the invention of Claim 23.

FIG. 15 is a flowchart illustrating the operating procedure of the example of FIG. 14.

FIG. 16 is a block diagram illustrating the functional configuration of an example of the invention of Claim 20.

FIG. 17 is a flowchart illustrating the operating procedure of the example of FIG. 16.

FIG. 18A is a drawing illustrating an example of the application of noise and signals in a simulation experiment to which the invention of Claim 23 is applied, and FIG. 18B is a drawing illustrating the experimental conditions thereof.

FIG. 19 is a drawing illustrating the experimental results of FIG. 18.

FIG. 20 is a flowchart for explaining a conventional technique.

FIG. 21 is a drawing illustrating a comparative example of a time waveform of air conduction sound and bone conduction sound generated in a laboratory environment with noise (pink noise) of 90 dBA.

(FIG. 1)

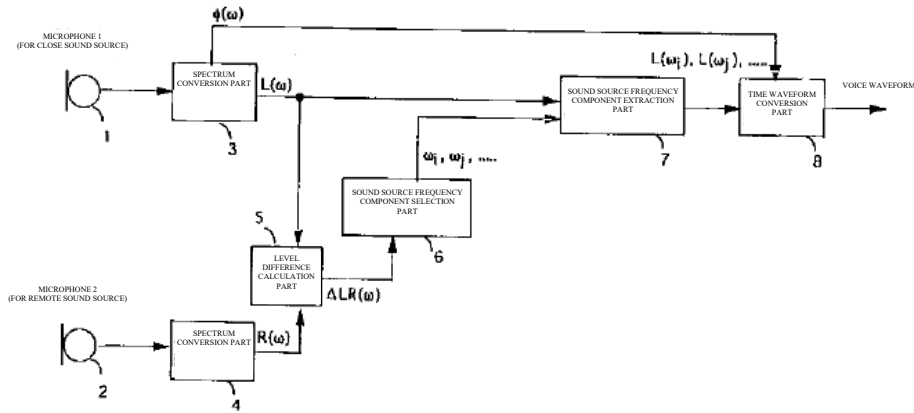


FIG. 1

(FIG. 2)

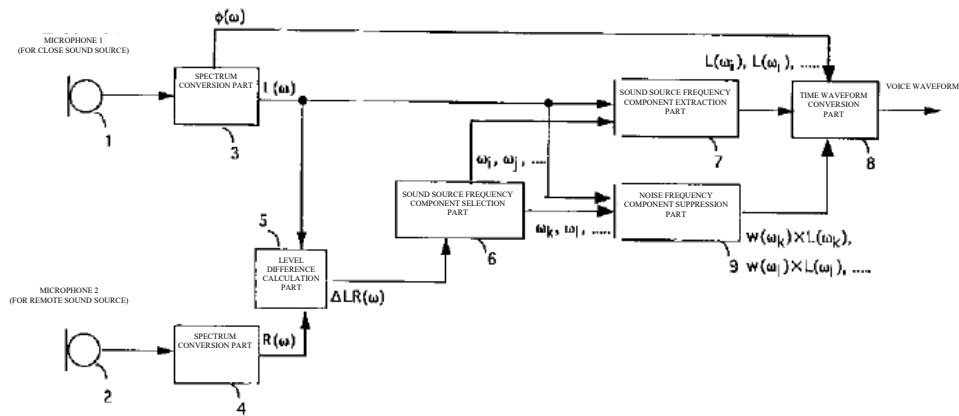


FIG. 2

(FIG. 3)

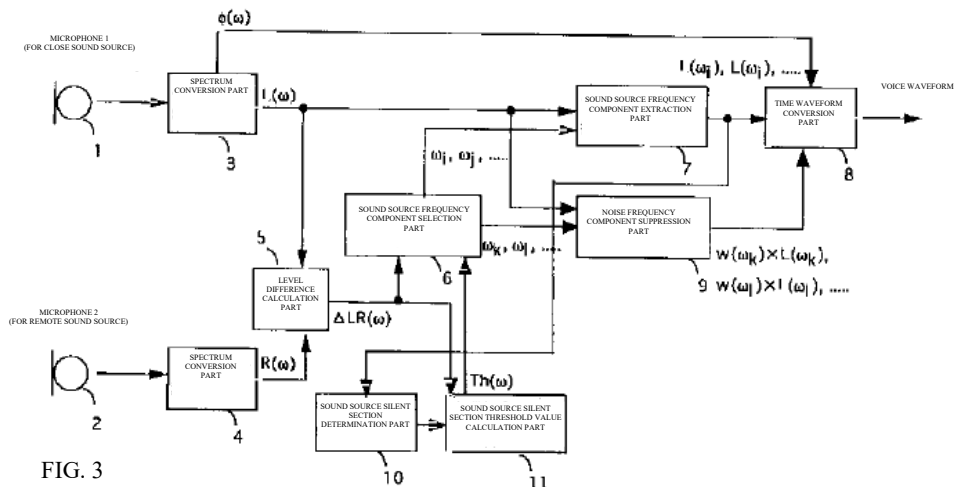


FIG. 3

(FIG. 4)

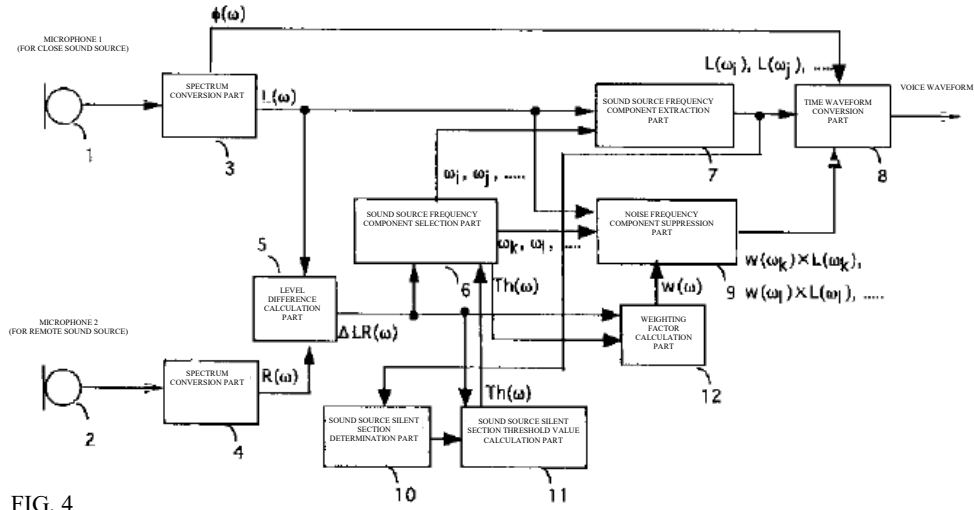


FIG. 4

(FIG. 5)

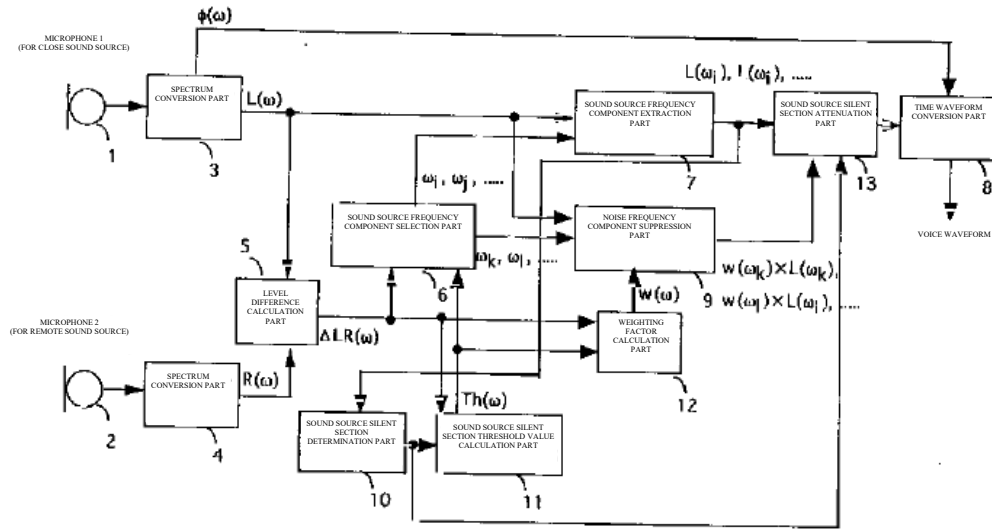


FIG. 5

(FIG. 6)

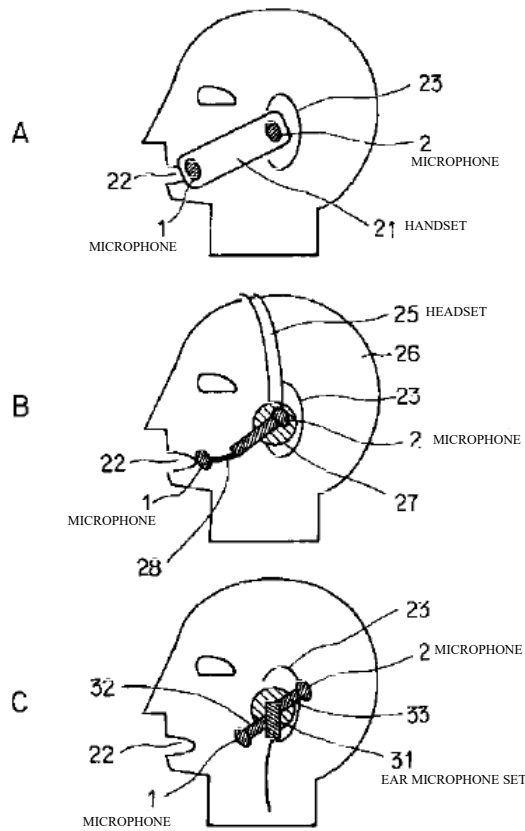


FIG. 6

(FIG. 7)

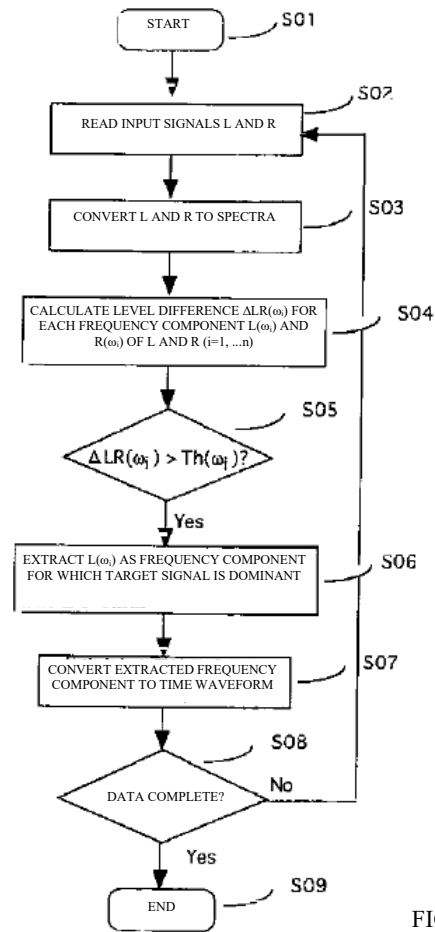


FIG. 7

(FIG. 12)

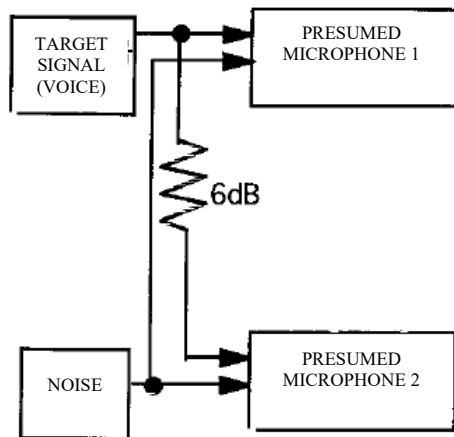


FIG. 12

(FIG. 19)

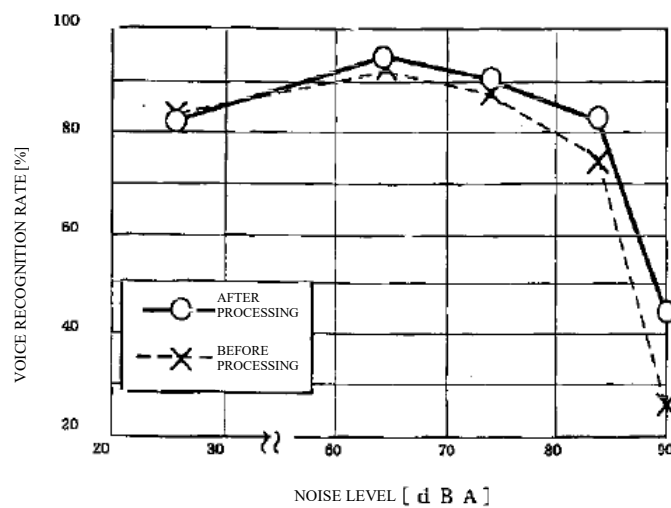


FIG. 19

(FIG. 8)

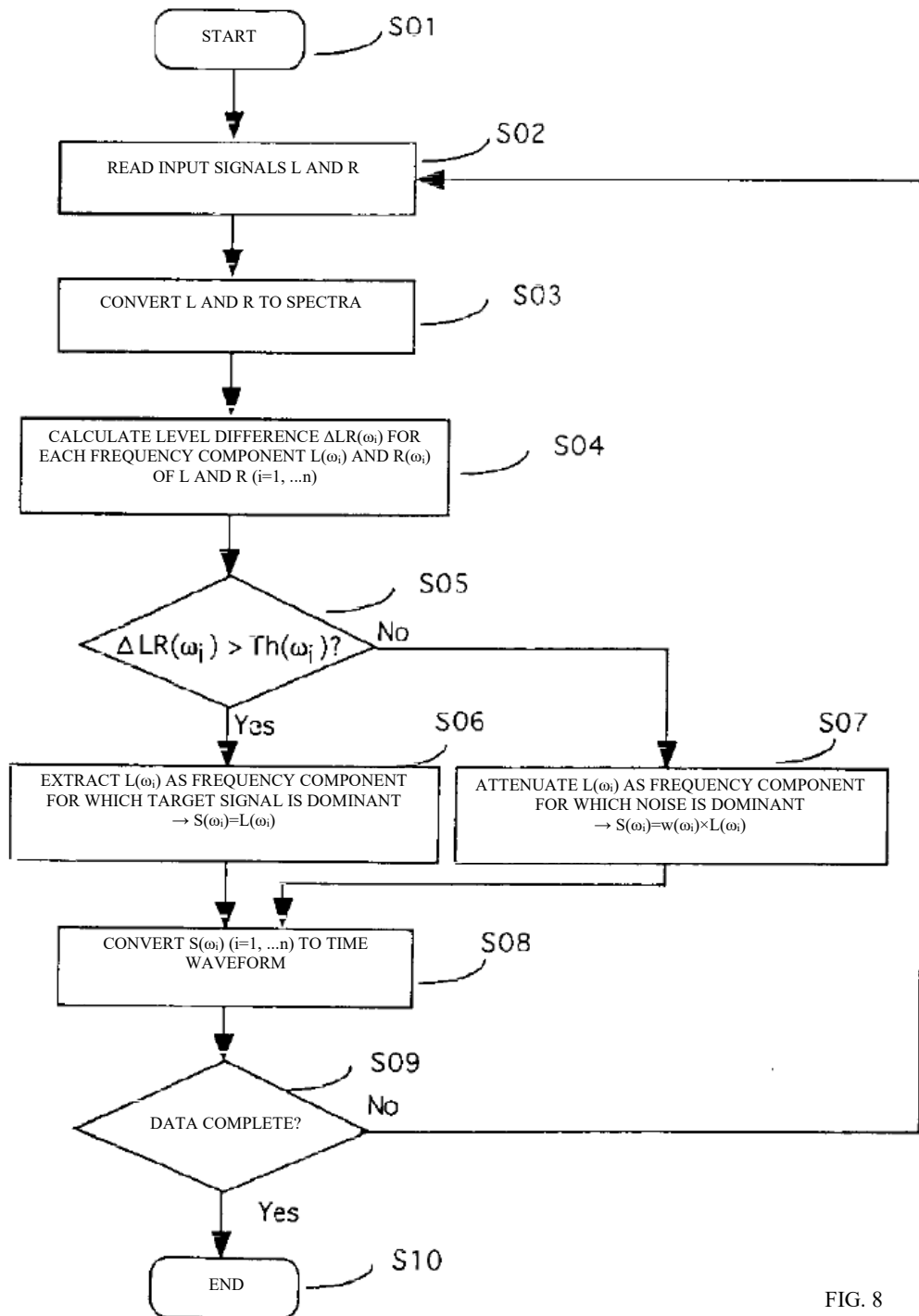


FIG. 8

(FIG. 9)

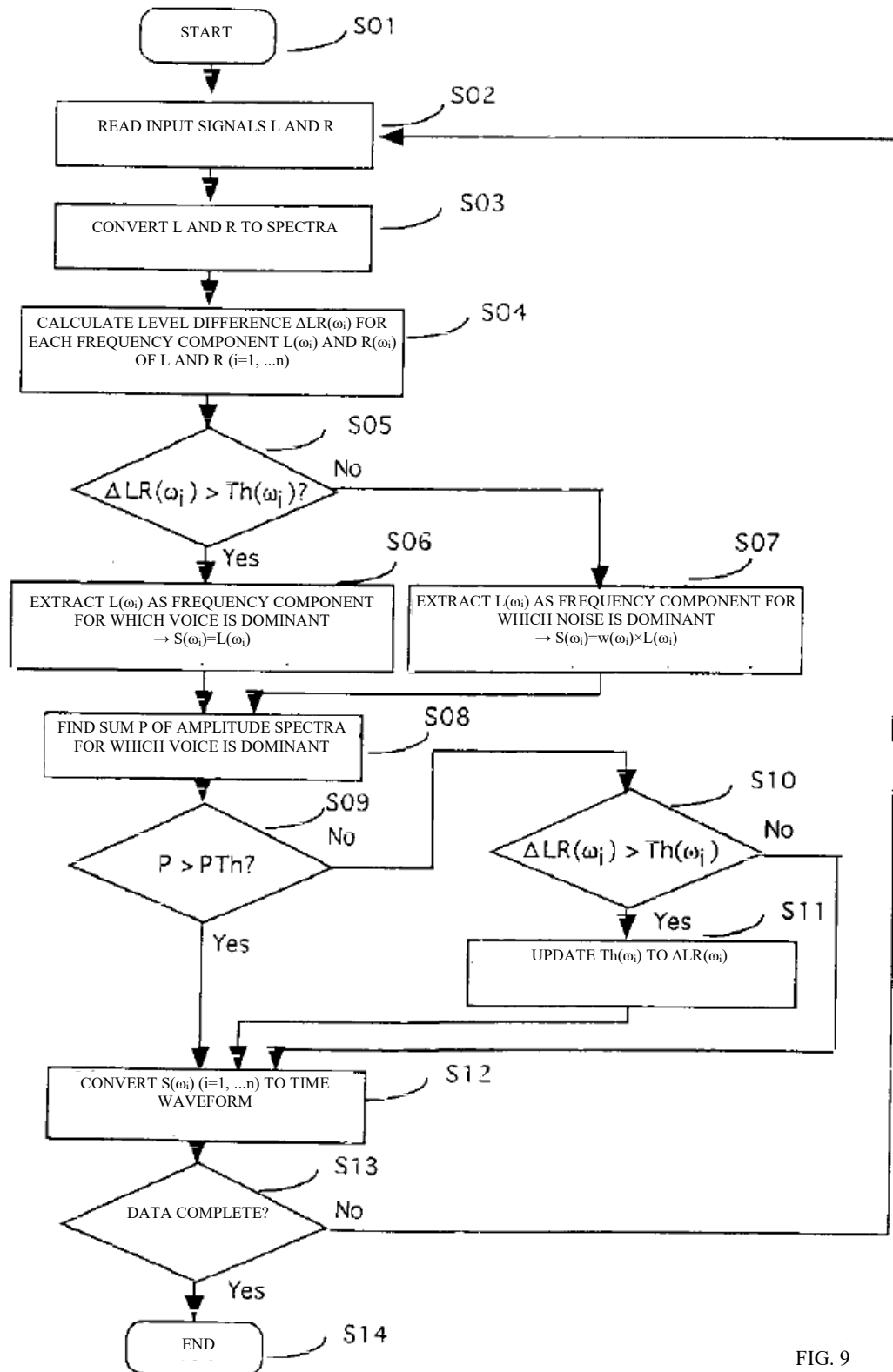


FIG. 9

(FIG. 10)

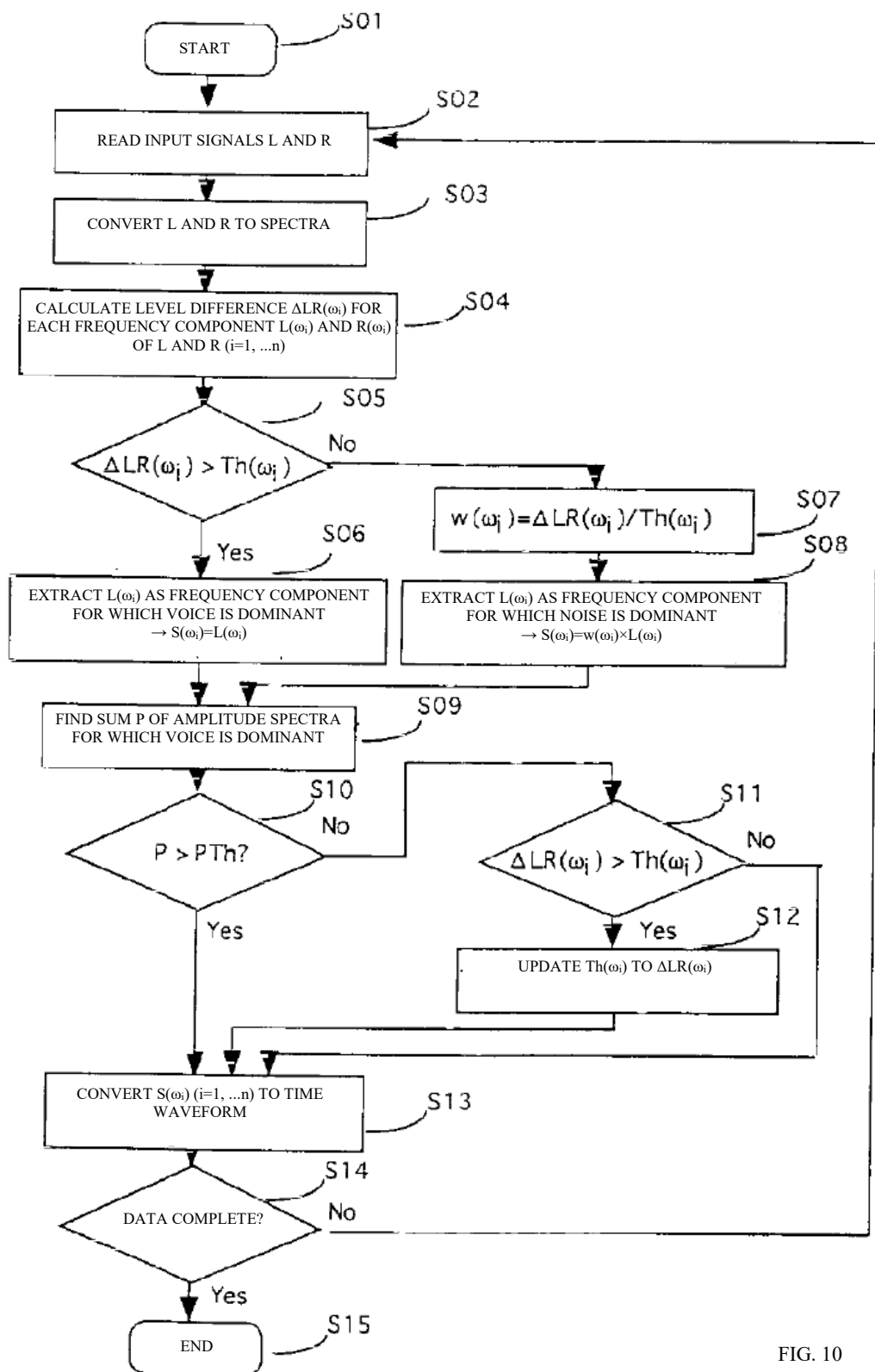


FIG. 10

(FIG. 11)

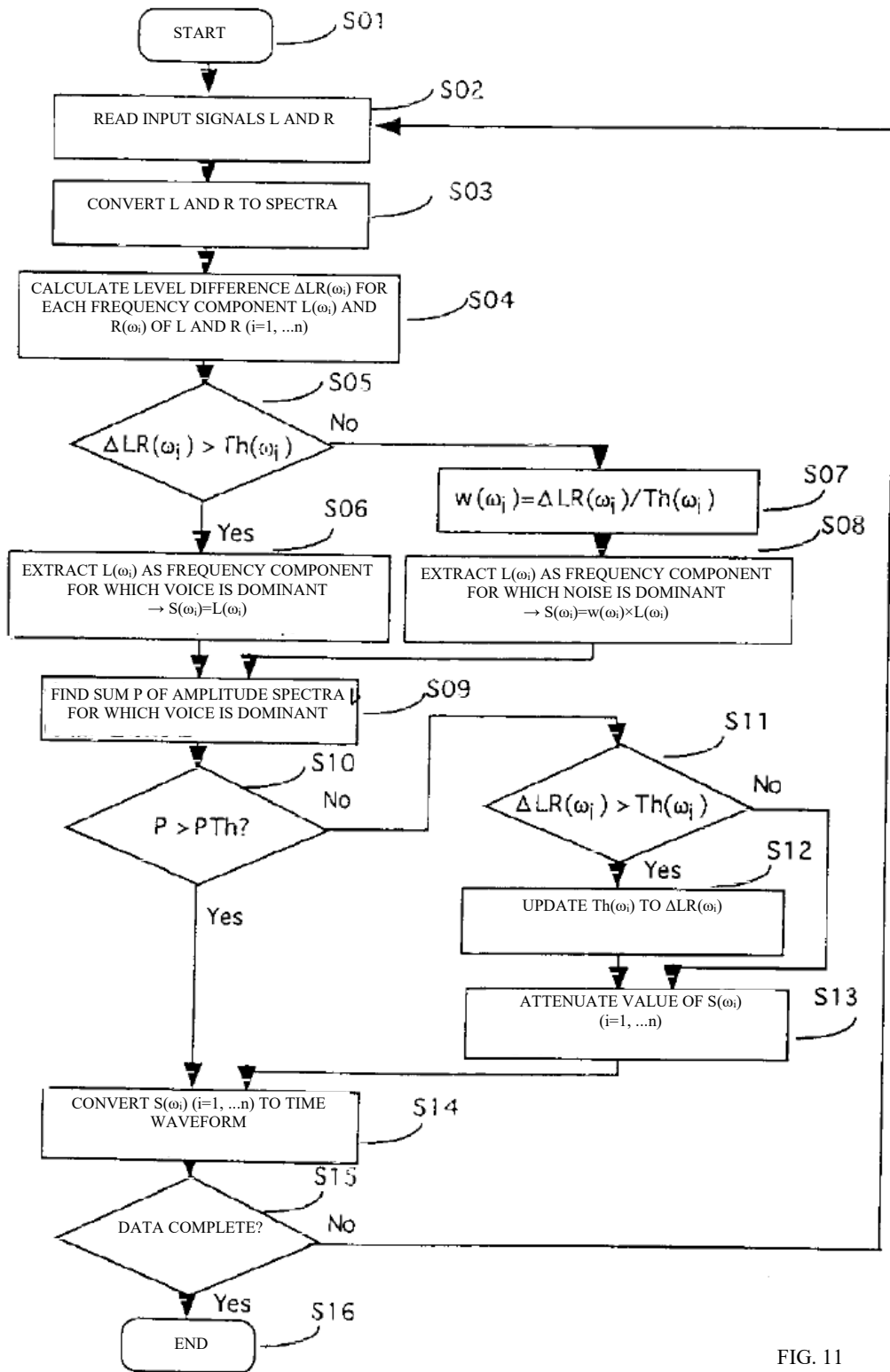


FIG. 11

(FIG. 13)

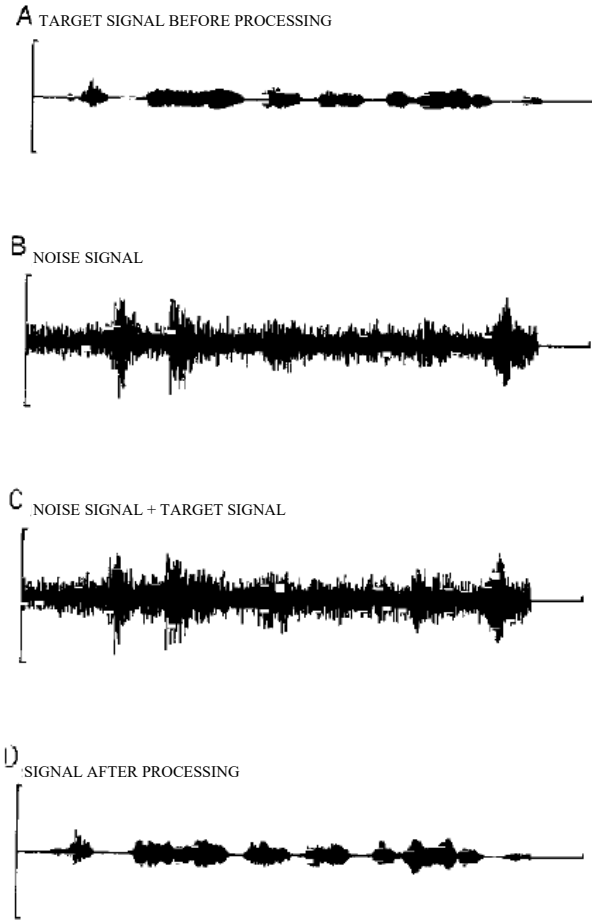


FIG. 13

(FIG. 15)

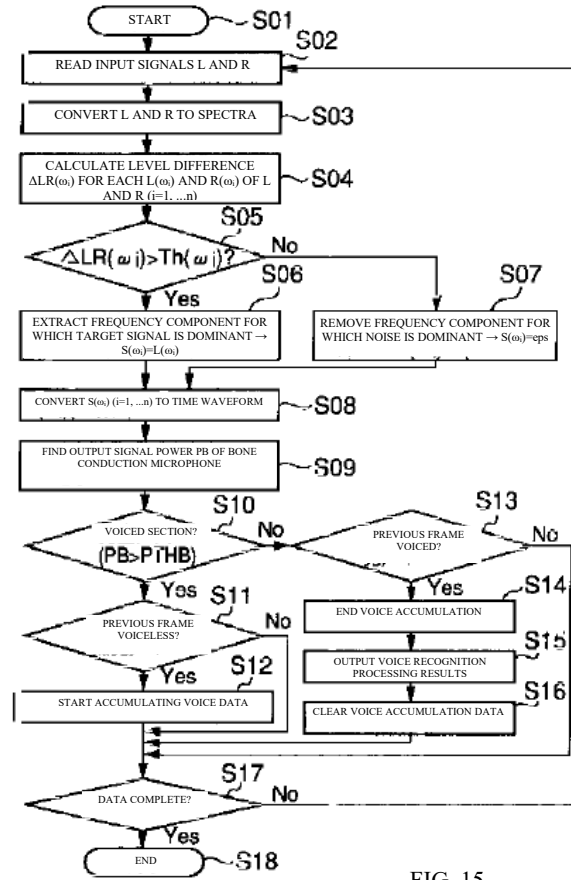


FIG. 15

(FIG. 14)

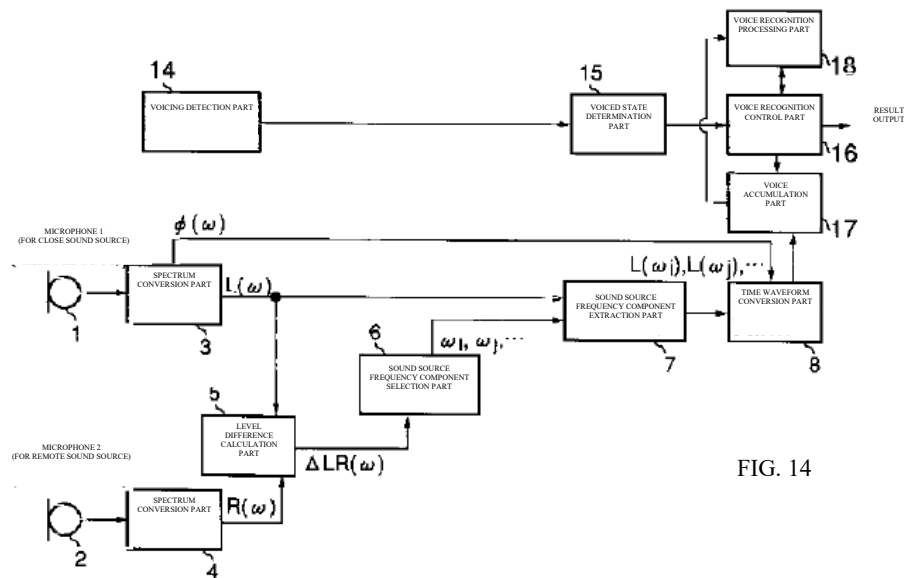


FIG. 14

(FIG. 16)

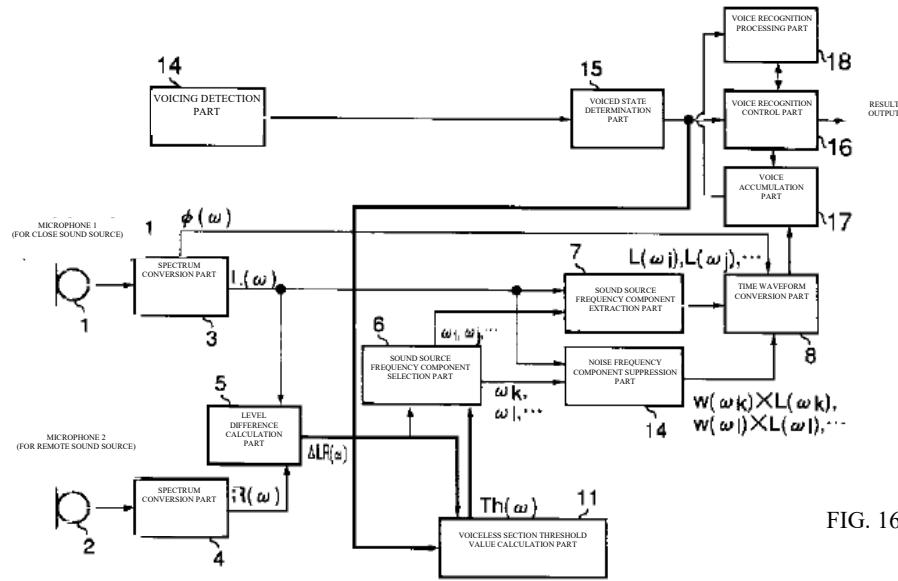


FIG. 16

(FIG. 17)

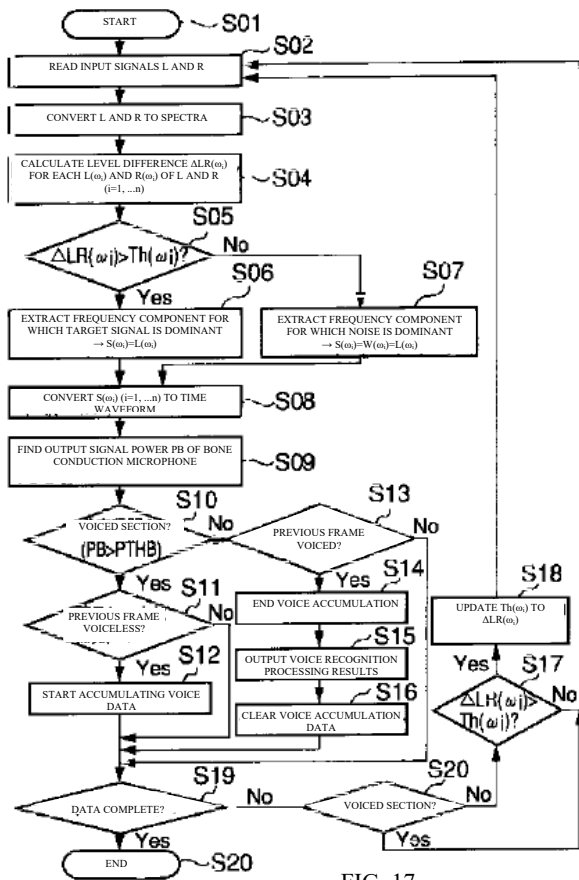


FIG. 17

(FIG. 18)

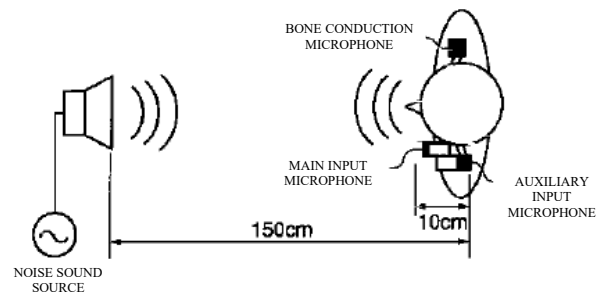


FIG. 18A

VOICING CONDITIONS	- SUBJECTS: 3 MALES - JAPANESE CITY NAMES: 20 TERMS
NOISE CONDITIONS	- NOISE SOUND SOURCE: PINK NOISE - NOISE LEVEL: 26 (WHEN NOISE SOUND SOURCE IS STOPPED), 64, 74, 84, 90 dBA
VOICE RECOGNITION PROCESSING	- WORD VOICE RECOGNITION USING SIMPLE DP MATCHING^[1] - USING LPC CEPSTRUM ANALYSIS - NOISE LEVEL AT TIME OF REGISTRATION: 64 dBA

^[1] See Nakadai, Sugimoto, Sakurai, and Nishino, Journal of The Institute of Image Information and Television Engineers, D-675, 3, 1996

FIG. 18B

(FIG. 20)

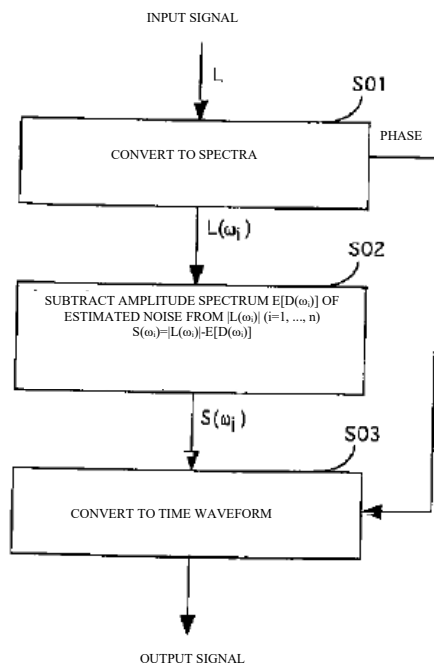


FIG. 20

(FIG. 21)

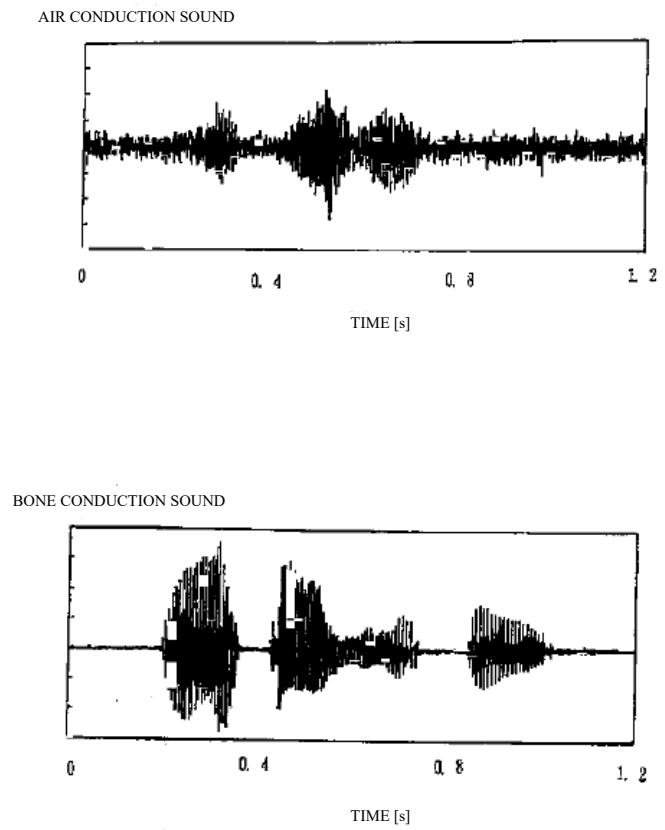


FIG. 21