United States Patent

Beninghof, Jr.

[15] **3,639,691**

[45] **Feb. 1, 1972**

[54] CHARACTERIZING AUDIO SIGNALS

[72] Inventor: **William J. Beninghof, Jr.**, Roslindale, Mass.

[73] Assignee: **Perception Technology Corporation**, Winchester, Mass.

[22] Filed: **May 9, 1969**

[21] Appl. No.: **823,372**

[52] **U.S. Cl.** .....179/1 VS

[51] **Int. Cl.** G10l 1/12

[58] **Field of Search**.....179/1 VS, 15.55 TC; 84/464; 324/77

[56] **References Cited**

UNITED STATES PATENTS

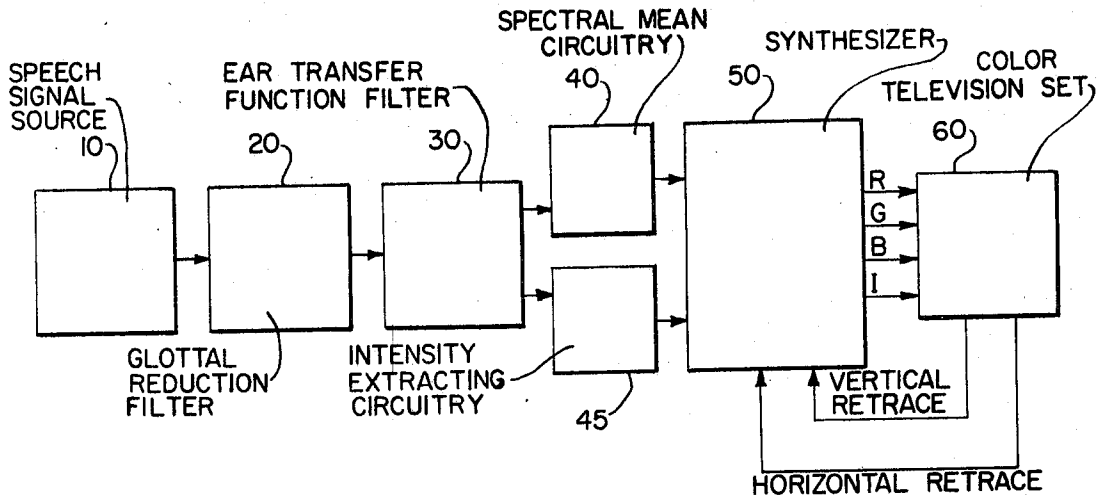
2,804,500	8/1957	Giacoletto	179/1 VIS
3,045,181	7/1962	Taffe et al.	179/1 VIS X
3,163,077	12/1964	Shank	179/1 VIS X
3,304,369	2/1967	Dreyfus	179/1 VIS
2,476,445	7/1949	Lacy	179/1
3,104,284	9/1963	French et al.	179/15.55

Primary Examiner—Kathleen H. Claffy
Assistant Examiner—Horst F. Brauner
Attorney—Wolf, Greenfield, Hieken & Sacks

[57] **ABSTRACT**

The loudness, spectral mean and spectral spread of speech signals are represented in the visual domain similar to brightness, hue and saturation of a color, respectively. The above parameters of a speech signal are extracted, and, by various operations, adapted for use in, and/or with other systems. As in color, the values of these parameters are defined relative to reference frames such that the parameters so extracted are to a large degree insensitive to extraneous ambient noises, speaker differences and overall (wideband) filterings.

13 Claims, 4 Drawing Figures



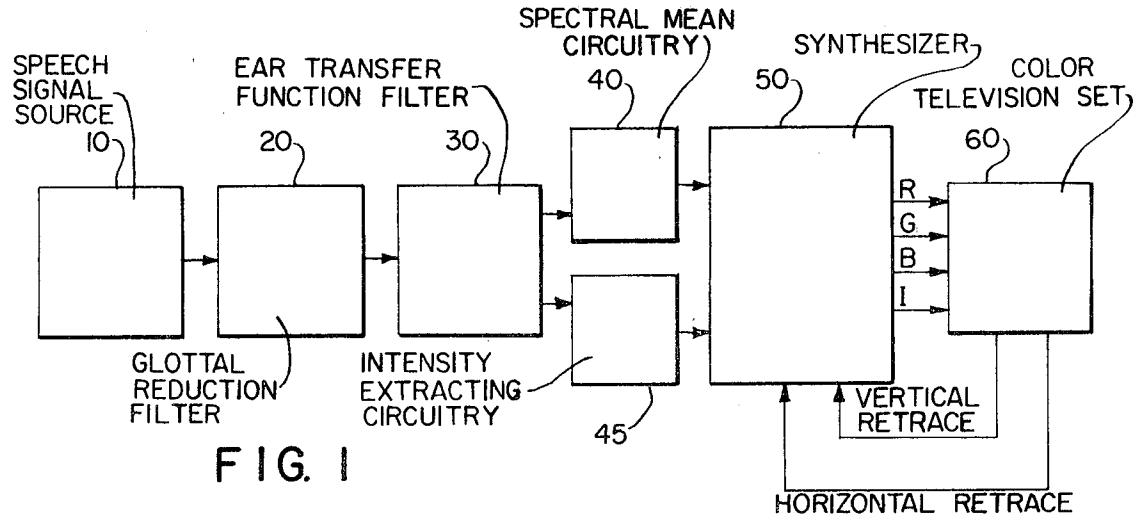


FIG. 1

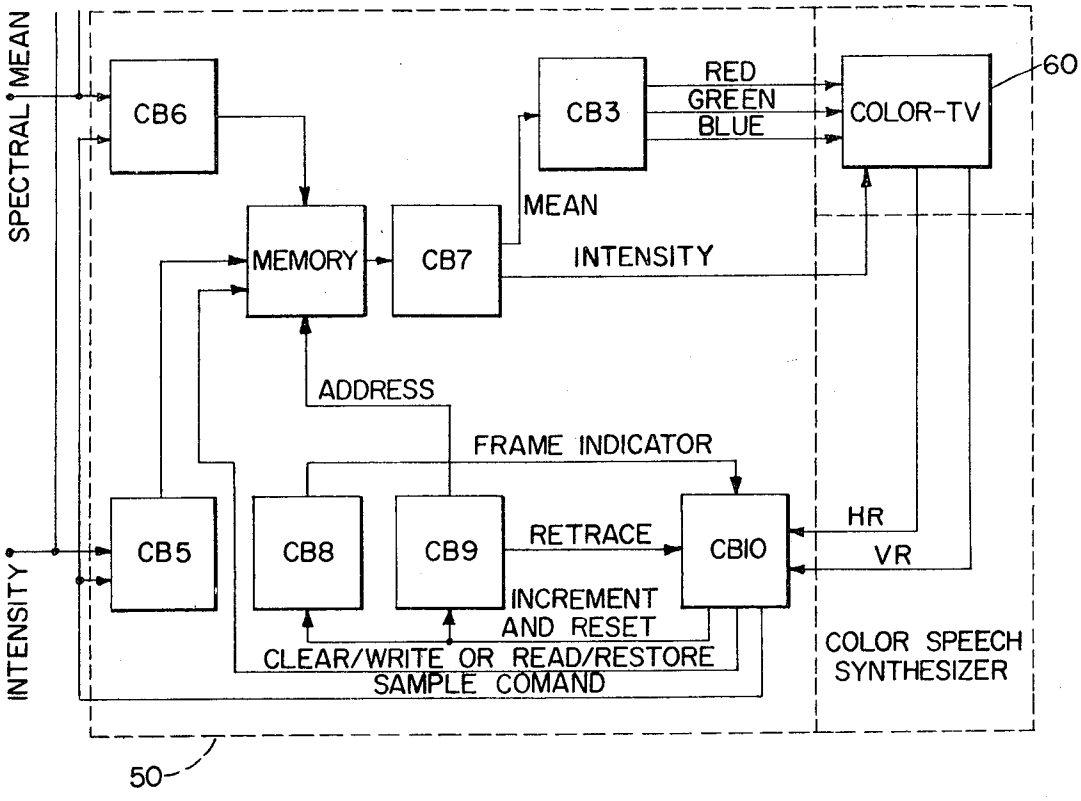


FIG. 2

INVENTOR
WILLIAM J. BENINGHOF JR.
BY
Wolfe, Greenfield, Hicken & Starks
ATTORNEYS

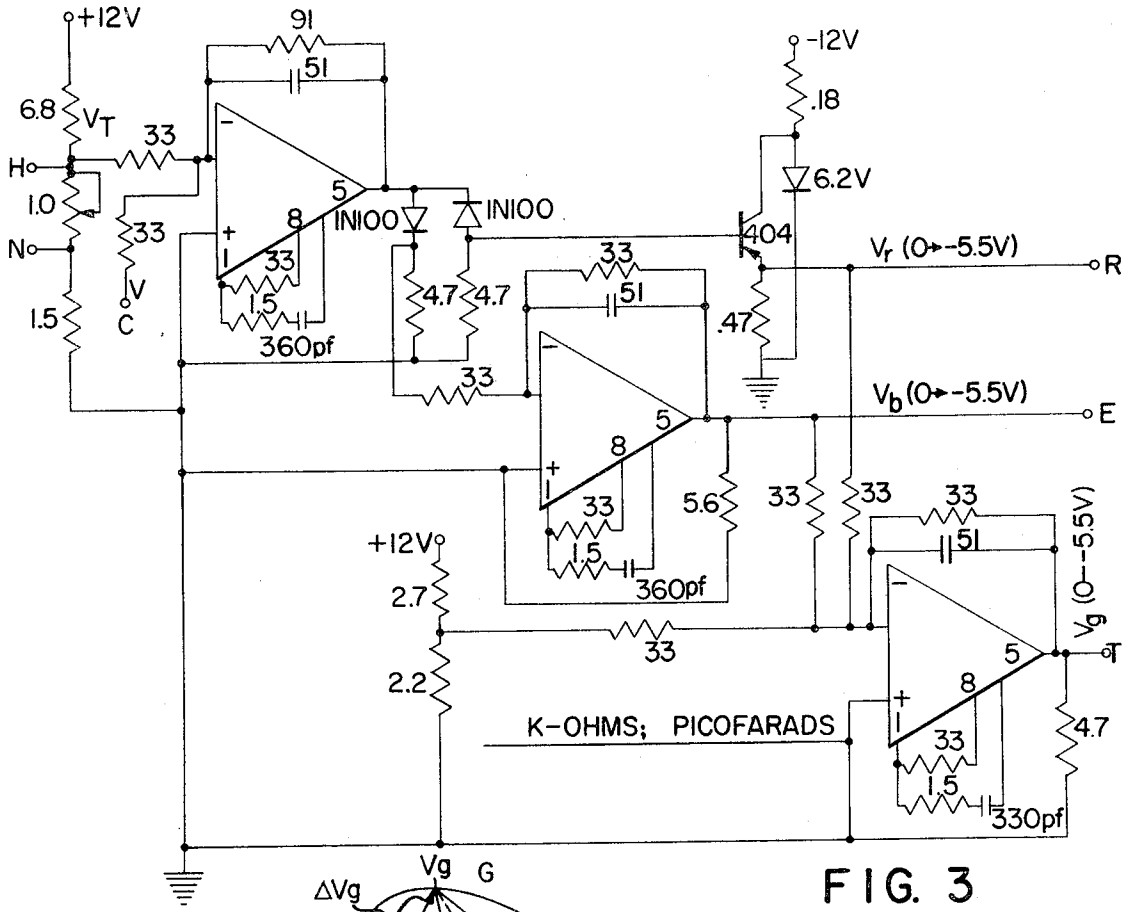


FIG. 3

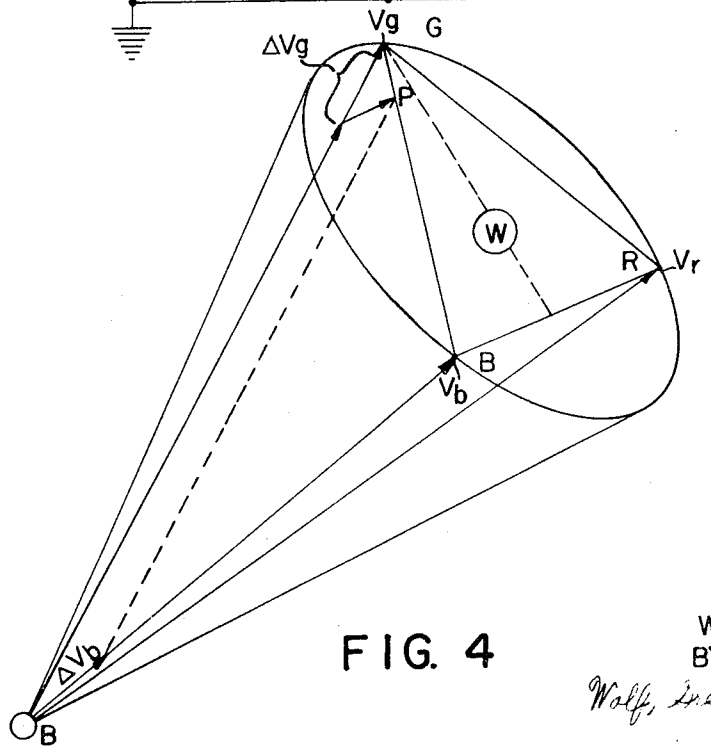


FIG. 4

INVENTOR
WILLIAM J. BENINGHOF JR.
BY
Wolfe, Greenfield, Hieken & Hocks
ATTORNEYS

CHARACTERIZING AUDIO SIGNALS

BACKGROUND OF THE INVENTION

The present invention relates in general to characterization of an audio signal and more particularly concerns a novel means and method of characterizing an audio signal as color. This application includes subject matter described in a thesis submitted in May, 1968 to the Department of Electrical Engineering of Northeastern University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the field of Speech Communications by William J. Beninghof, Jr. entitled A FUNCTIONAL ANALOGY BETWEEN SPEECH AND COLOR PERCEPTION AND ITS IMPLEMENTATION FOR SENSORY REPLACEMENT available in the Northeastern University library prior to May 9, 1969.

In some well-known methods of characterizing audio signals, the signal is displayed in a rectangular coordinate system. The time duration frequency and energy of the signal are represented by the abscissa (x), the ordinate (y) and the darkness (z), respectively. These displays, called spectrograms, reveal resonances (called formants) in the vocal track and have been considered to be an important tool for speech research. However, the spectrogram display lacks characteristics intelligible to the human sensory system. The viewer must estimate the frequencies of the first two formants and make a logical judgment of the combined frequencies to receive information. The theoretical foundation (the analogy between speech and color) for this invention results in a visual representation of speech which is most efficient. The viewer receives the information as a single sensory impression. He does not have to do any decoding in order to interpret the pattern.

It is an important object of the invention to provide improved methods and means for characterizing audio signals.

It is another object of the invention to provide methods and means for characterizing an audio signal which utilize the functional model of perception which is supported by the psychological response characteristics and physiological considerations of both human auditory and visual systems.

It is another object of the invention to provide methods and means for characterizing an audible signal wherein the extracted parameters may be used as inputs for systems for communication, recognition, bandwidth compression, control, data processing and speech training systems.

Another object of the invention is to provide methods and means for mapping an audio signal into a characterizing color display.

It is another object of the invention to provide methods and means for characterizing audio signals using a storage display which permits the viewer to use temporal cues, contextual constraints and transitional information normally available to a listener as a result of a more highly developed temporal memory.

Another object of the invention is to provide methods and means for characterizing an audio signal which is capable of immediate feedback so that the relation between the audio signal and the visual display can be obtained in the process of relating the audio signal to the visual display variables.

It is another object of the invention to provide methods and means for characterizing an audio signal providing for a visual or color display of speech, recognition of speech and bandwidth reduction wherein such means and methods may be adapted to systems such as deaf trainers, language trainers, vocabulary recognizers, phoneme recognizers, and communication systems with efficient and/or reduced bandwidths.

It is another object of the invention to provide methods and means for characterizing an audio signal which facilitates the recognition of audio signals, particularly by deaf humans.

It is a further object of the invention to provide a method and means for characterizing an audio signal which is suitable for implementation in a two-way communications medium so that visual information representative of the audio signal may be perceived by a deaf human.

It is another object of the invention to present, extract and utilize the parameters of speech in a way relatively insensitive to extraneous noises, speaker differences, and (overall) wideband filtering.

It is another object of the invention to use the extracted signals which are relatively independent of extraneous ambient noises, speaker differences and wideband filtering for recognition of speech sounds, phonemes and words.

It is another object of the invention to use the extracted signals which are relatively independent of extraneous ambient noises, speaker differences and wideband filtering for the purpose of bandwidth reduction, since control variables to define the perceptual reference frame require less information than the variations caused by the perceptual shifts themselves.

It is another object of the invention to use the extracted signals for design and operation of efficient communication systems.

It is another object of the invention to provide methods and means for characterizing an audio signal which permit the use of context and syntax in identifying the particular utterance of the audio signal in the visual display.

Another object of the invention is to achieve one or more of the preceding objects while keeping costs relatively low.

SUMMARY OF THE INVENTION

According to the invention there is means for extracting parameters which are perceptually significant and efficiently representative of speech sounds from speech-representative electricals signals. Means responsive to signals characteristic of these parameters produce intelligible speech displays and may comprise means for effective speech recognition and for use in effective communications systems. A feature of the invention is that the parameters form a closed perceptual space similar to the parameters for representing colors helpful in providing intelligence representative signals from the speech signals relatively independent of the characteristics of the speaker and the effects of ambient noise. Stated in other words, characteristics of a sound may be identified as points in a closed curve in a manner similar to the identification of the saturation and hue of colors as points in a closed curve.

The invention includes several embodiments for carrying out the stated objects. One embodiment may characterize audio signals in a visual display by means of a horizontal sequence of vertical strips of colors which correspond to the temporal sequence of audio signals. Another embodiment of the invention may characterize audio signals by means of visually analogous electrical parameters suitable for an input signal for control, communication or data processing systems. In yet another embodiment, the pitch is extracted and embodied in the display.

In the first embodiment of the invention a speech analyzer extracts two slowly varying parameters from the audio signal, such as the intensity and the spectral mean. The audio signal is transformed by a filter circuit to eliminate glottal contribution in the spectral distribution so that the personal attributes of speech, such as pitch and quality, which are related to the identification of the speaker, will not affect the display. The transformed signal is further transformed in the informational domain by electronic circuitry having a transfer function closely simulating that of the ear. The spectral mean and the intensity of the audio signal are then separated. The respective signals are then converted to digital form and placed in the memory. The memory is synchronized with the rest of the system by clocking and counting circuitry which allows the data to be time-compressed and converted for display on a color cathode ray tube, as for example, color television tube. The signal representative of the spectral mean is converted to three signals, subject to a constant brightness constraint and thus may be regarded as normalized with respect to brightness, which may be impressed on the grids of the color amplifier so as to control the three color guns of a modified color television set. The time-compressed signal representa-

tive of the intensity of the audio signal manifests itself as a binary voltage and interconnects with the cathode-ray tube to darken the particular portion of the display when there is no utterance. The usual scanning circuitry of the television is used to provide the trace and retrace for the visual display and provides the command and clocking signals for the entire system. Thus, the audio signal is represented in visual manner by a real-time display of a horizontal sequence of vertical strips of colors which correspond to the temporal sequence of the audio signal. At any given point in time the visual display represents the last 4 seconds of an audio signal. As time progresses, the pattern moves from right to left in a ticker-tape manner. Moreover, to facilitate recognition of the patterns the display may be stopped and the colors representative of the audio signals may be examined for an indefinite period.

In other embodiments of the invention, the signals representative of the parameters of the audio signals may be placed in the memory as above. The information in the memory may be extracted on command and transferred as a stimulus to control apparatus, communication circuitry or may be placed in another memory for indefinite storage. The pitch of the audio signal, by which vocal inflections may be distinguished, may be represented in the first embodiment by a flicker of the brightness on the face of the cathode ray tube.

Numerous other features, objects and advantages of the present invention will be more clearly understood when considered in conjunction with the accompanying drawings in which:

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. **1** is a block diagram illustrating a specific embodiment of the invention and further showing the signal flow through the apparatus;

FIG. **2** is a block diagram of the color-speech synthesizer according to the invention;

FIG. **3** is a schematic circuit diagram of transfer circuitry CB_3 of the color-speech synthesizer; and

FIG. **4** is a diagram of the color cone wherein the rays of the color cone represent the chromaticity of the three phosphors.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

With reference now to the drawings and more particularly to FIG. **1** thereof there is shown a block diagram illustrating a specific embodiment of the invention and the signal flow through the system. The source of audio signal **10** couples to and supplies the input signal for the apparatus. Filter **20** modifies the audio signal and eliminates the glottal contribution to the spectral distribution which has no counterpart in the chromatic sensation and whose presence would otherwise make the determination of spectral mean less stable.

The audio signal is further modified by filter **30** which closely approximates the transfer function of the ear (the ratio of the velocity of the basilar membrane to pressure at the pinna). The output of filter **30** permits a stable determination of the spectral mean and intensity.

Spectral mean circuitry **40** measures the short-time average zero-crossing density of the output signal of filter **30** and yields an output signal of filter **30** and yields an output signal proportional to the mean of the frequency distribution of its input signal. The output signal of the filter **30** also provides the input signal for intensity extracting circuitry **45** which is designed for dual operation.

The intensity portion of the signal may be transformed in a dual manner: (1) the output signal may be linearly related to the energy of the input signal and (2) the output signal may be binary, i.e., that is, one of two levels depending upon whether or not the energy of the input signal exceeds the certain threshold level. The output signals of spectral mean circuitry **40** and intensity extracting circuitry **45** provide the two slowly varying parameters which are the input for the color-speech synthesizer **50**.

Within the synthesizer **50** the two signals are sampled and stored in a digital memory. These digital signals are then transferred into a memory which is sampled and stored at the particular sampling rate. The output of the color-speech synthesizer **50** connects to the visual display, as for example, a color television set **60**.

The scanning portion of the color television set **60** feeds back to the synthesizer **50** and provides the timing reference for the entire system. The data within the memory is time-compressed due to the difference between the storing rate and the reading rate achieved from the horizontal retrace pulse of the television set. The information is then read out through digital-to-analog converters which convert the digital information into analog signals. The intensity signal is then connected to the grid of the video amplifier of the color television set so as to provide intensity information in a binary manner that is, display or nondisplay.

The time-compressed spectral means signal which controls the chromaticity of the display without changing the brightness is converted into three signals subject to a constant brightness constraint. These three signals provide the input signal for the color television display which consists of three phosphors of fixed chromaticity. In the usual manner of display, the color television rests on its left side so that the bottom of the television set becomes the right side when used in this display mode. The information is displayed as a horizontal sequence of vertical strips of colors which correspond to the temporal sequence of audio signals. For an exposition of the mathematics involved and the development of the analogy between physical stimuli received by the eye and ear, reference is made to the thesis mentioned above. That thesis sets forth theoretical considerations in section IV and relevant considerations by which Hermite functions may be used as a basis for forming a closed perceptual space.

FIG. **2** is a block diagram of color speech synthesizer **50** and further shows interconnections with color television set **60**. The output signals from the spectral mean circuitry **40** and the intensity extracting circuitry **45** provide the input signals for color speech synthesizer **50**. These signals are sampled every **15** msec. (the time for one television field), and converted into digital form by the analog-to-digital converters CB_6 and CB_5, respectively.

Converters CB_6 and CB_5 are serial analog-to-digital converters which count the series of clock pulses until the number stored in the counters represents the amplitude of the input signal waveform, at which time further counting is inhibited. The count remains fixed until it is time to transfer it into the memory after which counting is reset.

Provision is also made to block counting when full counting is reached regardless of input amplitude. This results in a small error between the digital and analog form of the data when the sample amplitude is larger than the capacity of the counter, but it prevents the gross error that would result if the counter cycled back to a small digital number.

The clock for both analog-to-digital converters is included in converter CB_6 and may be started by a "sample command" which permits it to generate a number of pulses greater than the maximum possible count. The digital data is then transferred from converter CB_5 and CB_6 to the memory at time of "clear or write command" which occurs once in every television field (**15** msec.). **262** words of data may be stored in the memory in any instant. Each word then corresponds to one of the **262** horizontal traces in the television field, and each may be read out of the memory on the "read or restore command" synchronously with the horizontal traces from the television set **60**.

The sampling and storing of the data at the field rate (15 msec.) and reading memory at the horizontal trace rate (60 msec.) provides the time compression for the store display. By incrementing the address register CB_9 an extra count once each field, the stored display may move from bottom to the top of the color television set. The display will then correspond to the most recent 4 seconds of an audio signal.

The data are read out of the memory through converter CB_7. The time-compressed intensity-representing signal is characterized by a binary voltage which drives the cathode-ray tube off when there is no utterance. Since this cathode-ray tube display consists of three phosphors of fixed chromaticity, the voltage corresponding to the spectral mean of the speech should control the chromaticity of this display without changing the brightness. Thus the effective chromaticity can be changed only by changing the brightness of the three phosphors. Transfer circuitry CB_3 supplies a good approximation to a constant brightness display by obtaining a converted set of three signals representative of the time-compressed spectral-mean signal which controls the intensity of the electron beam exciting the respective phosphors.

The timing of control pulses for the color-speech synthesizer **50** is derived from the horizontal retrace pulse of the television set **60**. These pulses are channeled through timing and control circuitry CB_{10} and counted by a standard counter CB_8 which provides indication of the sampling time of the speech data and of the beginning of a field of the television set **60**. The address register CB_9 also comprises a standard counter which provides the address of the word of memory being processed and provides for the ticker-tape effect in the visual display. Circuitry for shaping the pulse and timing is also included within address register CB_9.

The shaping and timing of the horizontal retrace (HR) pulse from television set **60** is done by the timing and control circuit CB_{10}. The input signals to the timing and control circuit CB_{10} are the vertical retrace (VR) pulse, the horizontal retrace pulse (HR) and the sample time indication from the trace counter CB_8. The latter is simply converted to a much wider pulse which "enables" the clock of the analog-to-digital converters CB_6 and CB_5. The vertical retrace (VR) pulse is effective only at turn-on time. It resets both the trace counter CB_8 and the address register CB_9 at the time of the horizontal trace which occurs near the bottom of the screen of television set **60**.

The result is that the bottommost horizontal trace represents the most recent speech sample. The sample is stored, and as it becomes older, it moves from the bottom to the top of the display finally disappearing from storage after about four seconds. The horizontal retrace pulse (HR) is basic to all other timing and control. It is shaped and delayed to provide pulses for incrementing and resetting the trace counter CB_8 and address register CB_9 and for providing read or restore and clear or right commands for the memory system.

FIG. **3** is a schematic circuit diagram of transfer circuitry CB_3 which converts the time-compressed spectral-mean representing voltage V to three voltages, V_b, V_g and V_r which control the intensity of the electron beam exciting the respective phosphors of the cathode-ray tube of television set **60**.

Transfer circuitry CB_3 can best be explained by reference to the color cone of FIG. **4** and by the criteria that the voltage corresponding to the spectral-mean of the audio signal should control the chromaticity of the display without changing the brightness. Since the cathode-ray tube display consists of three phosphors of fixed chromaticity, the effect of chromaticity can be changed only by changing the brightness of the three phosphors.

The rays of the color cone of FIG. **4** represent the chromaticity of the three phosphors. The distance along any ray indicates the brightness. A plane intersecting the three rays at points representing the brightness of the phosphors is an equal-brightness surface. Reducing the brightness of the green phosphor by an amount ΔV_g and increasing the brightness of the blue phosphor by an equal amount ΔV_b ideally yields a sensation of equal brightness with a chromaticity at a point P. In transfer circuitry CB_3 the equal brightness condition was imposed by the following two sets of constraints:

$$\text{A.}\ \left.\begin{array}{ll}(1)\ V_g+V_r=K\\(2)\ V_r=2(V+V_T)\\(3)\ V_b=0\end{array}\right\}\ |V|>V_T$$

$$\text{B.}\ \left.\begin{array}{ll}(4)\ V_g+V_b=K\\(5)\ V_b=-2(V+V_T)\\(6)\ V_r=0\end{array}\right\}\ |V|<V_T$$

The voltage criteria utilizing the above equations for transfer circuitry CB_3 are:

-5.5 volts $< V < 0$ volts

$V_T = +2.7$ volts and

$K = -5.5$ volts.

Geometrically and ideally these constraints amount to restricting the chromaticity and brightness to points along the two straight lines connecting B to G and G to R. The points B, and R are determined by the chromaticity of the phosphors and the brightness elicited by the maximum voltage $V = -5.5$ volts. In practice the voltages more nearly correspond to intensity than brightness and a decrease in brightness is observed for colors between the primaries. In addition the saturation as well as the hue changes as the chromaticity moves along the straight line locus. Although good yellows and aquas may not be produced by the transfer circuitry CB_3, the circuitry does provide adequately observable color differences for the purpose of characterizing sounds. More separation between audio signals in the color environment may be obtained by a more sophisticated set of constraints which would more nearly approximate the circumference of the color circle.

Having described the embodiments and the physical arrangement of the apparatus, it is appropriate to consider the results achieved in experimentation with the color-speech synthesizer. Audible signals in the form of numbers from 0 to 99 were recorded in random order, twice by a male speaker (recordings R_1 and R_2) and once by a female speaker (recording R_3). The recordings were played into the color-speech system and presented to three judges (J_1, J_2 J_3). The three judges were trained for 5 hours to identify the colors corresponding to recording R_1 after which they were tested on recording R_1 and achieved 98 percent, 96 percent and 83 percent correct identifications. Without any further training, or familiarization with recording R_2 and recording R_3, they were tested on these recordings and achieved 89 percent, 80 percent and 73 percent; and 65 percent, 63 percent and 43 percent correct identifications respectively.

Thus, an important feature of the invention is the real-time storage display and the "stop action" which enhance the learning process.

Another important feature of the invention is the perception of the color corresponding to an audio signal on a sensory level. As indicated above the viewer is not required to decode the display with information in excess of that which is needed and in a form which is not easily absorbed.

Another feature of the invention is the use of colors which serves to stimulate the learning process. This stimulation of the learning process would be of particular merit for training deaf children to articulate.

Another important feature of the invention is the utilization of storage display which is essential to the intelligibility of color speech because it provides temporal and contextual cues and facilitates the learning process in that it permits the sequence of colors corresponding to an audio signal to be stopped and studied indefinitely.

Another important feature of the invention is the adaptability of the system for use with control, or data processing apparatus. The invention may be connected, with compatible apparatus, to one or all of the above, thereby allowing a machine to recognize informational aspects of audio signals.

Another important feature of the invention revealed herein is the suitability and efficiency of the invention for machine recognition of speech. The parameters based upon the perceptual analogy between speech and color may be extracted and represented relative to a perceptual reference frame. These parameters and their relationship to the perceptual reference frame (which is time dependent) will permit efficient and ac-

curate recognition of speech. Such a recognition system is particularly useful in coding for the ultimate in reduction of a channel capacity for speech transmission and speech control systems.

Another important feature of the invention is the adaption of human perceptual techniques to achieve machine recognition of audio signals. The visual display may be implemented by a set of filters responsive to the color display wherein the optical energy passing through a word-representing color filter as a result of the color display moving past the filter is a measure of the correlation between the utterance moving past and the standard filter. A set of standard filters representing a limited vocabulary would provide a means of machine recognition of speech.

Another important feature of the invention is the adaptation of human perceptual techniques to achieve bandwidth compression for speech communication. The parameters used to represent the speech which are based upon the analogy between speech and color perception may be extracted, transmitted and used to synthesize speech. The bandwidth required for transmitting these parameters may be further reduced by transmitting those parameters which are measured relative to the perceptual reference frame. The perceptual reference frame may be specified by very slowly varying information.

Other modifications and uses of and departures from the specific embodiments described herein may be practiced by those skilled in the art without departing from the inventive concepts. Consequently, the invention is to be construed as embracing each and every novel feature and novel combinations of features present in or possessed by the apparatus and techniques herein disclosed and are limited solely by the spirit and scope of the appended claims.

What is claimed is:

1. Apparatus for characterizing speech signal comprising,

extracting means for extracting selected characterizing components of said speech signal to provide a characterizing signal therefrom and including glottal filter means for reducing glottal contribution in the spectral distribution of said speech signal and spectral mean circuit means for providing a slowly varying mean frequency electrical signal representative of the mean of the frequency distribution at its input,

storage means coupled to said extracting means for storing and time-compressing said slowly varying mean frequency signal comprising said characterizing signal,

transfer circuit means coupled to said storage means and responsive to the time-compressed stored electrical signal for transforming said stored time-compressed electrical signal into a speech output signal characteristic of said speech signal.

2. Apparatus for characterizing a speech signal according to claim 1 wherein said first output signal may be represented by visually observable hue and said second output signal may be represented by visually observable brightness.

3. Apparatus for characterizing a speech signal in accordance with claim 1 and further comprising,

means responsive to at least said slowly varying mean frequency electrical signal for providing each segment of said speech output signal representable as a point in a closed two-dimensional space which point represents a portion of a speech signal characteristic of a phoneme substantially independent of intensity and the personal attributes of speech such as pitch and quality which are related to the identification of the speaker.

4. Apparatus for characterizing a speech signal in accordance with claim 3 and further comprising color display means responsive to said speech output signal for providing a sequence of strips of color of substantially constant brightness which correspond to the temporal sequence of speech phonemes in said speech signal and slowly move across said display whereby an observer may comprehend the speech represented by said speech signal from observing the sequence of traveling color strips.

5. Apparatus for characterizing a speech signal in accordance with claim 3 and further comprising converting means for converting said characterizing signal into digital form and for providing an input characterizing signal for said storage means.

6. Apparatus for characterizing a speech signal in accordance with claim 4 and further comprising,

converting means for converting said characterizing signal to digital form and for providing an input characterizing signal for said storage means,

said color display means comprising color television set including scanning portion means coupled to said storage means and said transfer circuit means for providing timing and clocking signals for said storage and transfer circuit means.

7. Apparatus for characterizing a speech signal in accordance with claim 1 and further comprising,

intensity extracting circuit means for providing an intensity signal representative of the intensity of said speech signal and comprising said characterizing signal.

8. Apparatus for characterizing a speech signal in accordance with claim 1 and further comprising ear transfer function filter means for transferring components of said speech signal to said spectral circuit means in accordance with a transfer characteristic similar to that of the human ear.

9. Apparatus for characterizing a speech signal comprising,

an input terminal for receiving a speech signal to be characterized,

means for processing speech signal received on said input terminal for providing a speech output signal corresponding to a point in a closed two-dimensional space which point represents a portion of said speech signal,

and means for normalizing at least said output signal with respect to a predetermined parameter so that said point represents a sound phoneme substantially independent of personal attributes of speech such as pitch and quality that are related to the identification of the speaker.

10. Apparatus for characterizing an audio signal in accordance with claim 9 and further comprising means for storing a sequence of said first and second output signals for time-compressing a sequence of said portions.

11. A method of characterizing signals which method includes the steps of,

transforming a speech signal into an electrical signal having characteristics similar to those received by the human sensory system and having less dependence on pitch variations, intensity variations and speaker differences,

extracting information carrying components of said speech signal from said electrical signal corresponding at least to the mean of the frequency distribution of said electrical signal,

storing and time-compressing portions of said electrical signal,

and converting portions of the stored time-compressed electrical signal to an output signal representative of said information carrying components.

12. A method of characterizing audio signals according to claim 11 and further comprising displaying said output signals representative of said information carrying components.

13. A method of characterizing audio signals according to claim 7 wherein said output signal is displayed as strips of hues representative of said spectral mean in sequence substantially according to their relative temporal origin.

* * * * *