

5 TITLE: SPEAKER IDENTIFICATION AND VERIFICATION SYSTEM

10 BACKGROUND OF THE INVENTION

1. Field of the Invention

The present invention relates to a speaker recognition system or similar apparatus which applies
15 adaptive weighting to components in each frame of speech for normalizing the spectrum of speech, thereby reducing channel effects.

2. Description of the Related Art

The objective of a speaker identification
20 system is to determine which speaker is present from an utterance. Alternatively, the objective of a speaker verification system is to verify the speaker's claimed identity from an utterance. Speaker identification and speaker verification systems can be defined in the
25 general category of speaker recognition.

It is known that typical telephone switching systems often route calls between the same starting and ending locations on different channels. A spectrum of speech determined on each of the channels can have a
30 different shape due to the effects of the channel. In addition, a spectrum of speech generated in a noisy

environment can have a different shape than a spectrum of speech generated by the same speaker in a quiet environment. Recognition of speech on different channels or in a noisy environment is therefore difficult because
5 of the variances in the spectrum of speech due to non-vocal tract components.

Conventional methods have attempted to normalize the spectrum of speech to correct for the spectral shape. U.S. Patent No. 5,001,761 describes a
10 device for normalizing speech around a certain frequency which has a noise effect. A spectrum of speech is divided at the predetermined frequency. A linear approximate line for each of the divided spectrum is determined and approximate lines are joined at the
15 predetermined frequency for normalizing the spectrum. This device has the drawback that each frame of speech is only normalized for the predetermined frequency having the noise effect and the frame of speech is not normalized for reducing non-vocal tract effects which can
20 occur over a range of frequencies in the spectrum.

U.S. Patent No. 4,926,488 describes a method for normalizing speech to enhance spoken input in order to account for noise which accompanies the speech signal. This method generates feature vectors of the speech. A
25 feature vector is normalized by an operator function which includes a number of parameters. A closest

prototype vector is determined for the normalized vector and the operator function is altered to move the normalized vector closer to the closest prototype. The altered operator vector is applied to the next feature
5 vector in the transforming thereof to a normalized vector. This patent has the limitation that it does not account for non-vocal tract effects which might occur over more than one frequency.

Speech has conventionally been modeled in a
10 manner that mimics the human vocal tract. Linear predictive coding (LPC) has been used for describing short segments of speech using parameters which can be transformed into a spectrum of positions (frequencies) and shapes (bandwidths) of peaks in the spectral envelope
15 of the speech segments. Cepstral coefficients represent the inverse Fourier transform of the logarithm of the power spectrum of a signal. Cepstral coefficients can be derived from the frequency spectrum or from linear predictive LP coefficients. Cepstral coefficients can be
20 used as dominant features for speaker recognition. Typically, twelve cepstral coefficients are formed for each frame of speech.

It has been found that a reduced set of cepstral coefficients can be used for synthesizing or
25 recognizing speech. U.S. Patent No. 5,165,008 describes a method for synthesizing speech in which five cepstral

coefficients are used for each segment of speaker independent data. The set of five cepstral coefficients is determined by linear predictive analysis in order to determine a coefficient weighting factor. The
5 coefficient weighting factor minimizes a non-squared prediction error of each element of a vector in the vocal tract resource space. The same coefficient weighting factors are applied to each frame of speech and do not account for non-vocal tract effects.

10 It is desirable to provide a speech recognition system in which the spectrum of speech is normalized to provide adaptive weighting of speech components for each frame of speech for improving the vocal tract features of the signal while reducing the non-vocal tract effects.

15 SUMMARY OF THE INVENTION

The method of the present invention utilizes the fact that there is a difference between speech components and non-vocal tract components in connection with the shape of a spectrum for the components with
20 respect to time. It has been found that non-vocal tract components, such as channel and noise components, have a bandwidth in the spectrum which is substantially larger than the bandwidth for the speech components. Speech intelligence is improved by attenuating the large
25 bandwidth components, while emphasizing the small bandwidth components related to speech. The improved

speech intelligence can be used in such products as high performance speaker recognition apparatus.

The method involves the analysis of an analog speech signal by converting the analog speech signal to digital form to produce successive frames of digital speech. The frames of digital speech are respectfully analyzed utilizing linear predictive analysis to extract a spectrum of speech and a set of speech parameters known as prediction coefficients. The prediction coefficients have a plurality of poles of an all pole filter characterizing components of the frames of speech. Components of the spectrum can be normalized to enhance the contribution of the salient components based on its associated bandwidth. Adaptive component weightings are applied to the components of the spectrum to enhance the components associated with speech and to attenuate the components associated with non-speech effects. Cepstral coefficients are determined based on the normalized spectrum to provide enhanced features of the speech signal. Improved classification is performed in a speaker recognition system based on the enhanced features.

Preferably, the speaker recognition system of the present invention can be used for verifying the identity of a person over a telephone system for credit card transactions, telephone billing card transactions

and gaining access to computer networks. In addition, the speaker recognition system can be used for voice activated locks for doors, voice activated car engines and voice activated computer systems.

5 The invention will be further understood by reference to the following drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a flow diagram of the system of the present invention during training of the system.

10 Fig. 2 is a flow diagram of the system of the present invention during evaluation.

Fig. 3 is a flow diagram of the method of the present invention for feature extraction and feature enhancement.

15 Fig. 4 is a graph of a prior art speech spectrum without adaptive component weight filtering.

Fig. 5 is a graph of the speech spectrum shown in Fig. 4 with adaptive component weight filtering.

20 Fig. 6A is a spectrum without adaptive component weight filtering.

Fig. 6B is a spectrum with adaptive component weight filtering.

Fig. 7 is a comparison of spectrum with and without adaptive component weight filtering.

25 Fig. 8 is a response of a moving average (FIR) filter for a transfer function $(1-0.9z^{-1})$

DETAILED DESCRIPTION OF THE INVENTION

During the course of the description, like numbers will be used to identify like elements according to the different figures which illustrate the invention.

5 Fig. 1 illustrates a flow diagram of speech recognition system 10 during training of the system. A speech training input signal is applied to an analog to digital converter 11 to provide successive frames of digital speech. Feature extraction module 12 receives
10 the frames of digital speech. Feature extraction module 12 obtains characteristic parameters of the frames of digital speech. For speaker recognition, the features extracted in feature extraction module 12 are unique to the speaker to allow for adequate speaker recognition.

15 Feature enhancement module 14 enhances the features extracted in feature extraction module 12. Feature enhancement module 14 can also reduce the number of extracted features to the dominant features required for speaker recognition. Classification is performed in
20 block 16 on the enhanced features. Preferably, classification can be performed with a conventional technique of vector quantization in order to generate a universal code book for each speaker. In the
25 alternative, classification can be performed by multilayer perceptions, neural networks, radial basis function networks and hidden Markov models. It will be

appreciated that other classification methods which are known in the art could be used with the teachings of the present invention.

In Fig. 2, speaker recognition system 10 is shown for speaker identification or verification. A speech evaluation input signal is digitized in analog to digital converter 11 and applied to feature extraction module 12. Enhanced features of the speech input signal are received at a template matching module 18. Template matching module 18 determines the closest match in the universal code book or typical classification system for either determining the speaker's identification or for verifying that the speaker has an entry in the universal code book.

Fig. 3 illustrates a flow diagram of a preferred embodiment for the implementation of feature extraction block 12 and feature enhancement block 14. A frame of speech $s(k)$, can be represented by a modulation model (MM). Modulation model (MM) includes parameters which are representative of the number, N , of amplitude modulated (AM) and frequency modulated (FM) components. The frame of speech can be represented by the following formula:

$$s(k) = \sum_{i=1}^N A_i(k) \cos(\phi_i(k)) + \eta(k),$$

(100)

wherein $A_i(k)$ is the amplitude modulation of the i^{th} component, $\phi_i(k)$ is the instantaneous phase of the i^{th} component, and $\eta(k)$ is the modeling error.

Amplitude modulation component $A_i(k)$ and instantaneous phase component $\phi_i(k)$ are typically narrow band signals. Linear prediction analysis can be used to determine the modulation functions over a time interval of one pitch period to obtain:

$$A_i(k) = |G_i| e^{-B_i k}$$

(102)

and

$$\phi_i(k) = \omega_i k + \theta_i$$

15

(104)

wherein G_i is the component gain, B_i is the bandwidth, ω_i is the center frequency and θ_i is the relative delay.

20

Speech signal $s(k)$ is applied to block 110 to obtain linear predictive coding (LPC) coefficients. A LP polynomial $A(z)$ for a speech signal can be defined by the following equation:

10

$$A(z) = 1 + \sum_{i=1}^P a_i z^{-i}$$

(106)

wherein a_i are linear prediction coefficients and P is the order of the coefficients.

In linear predictive coding analysis, the transfer function of the vocal tract can be modeled by a time varying all pole filter given by a P th order LP analysis defined by the following:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + \sum_{i=1}^P a_i z^{-i}}$$

10

(108)

The roots of $A(z)$ can be determined in block 112 by factoring the LP polynomial $A(z)$ in terms of its roots to obtain:

15

$$A(z) = \prod_{i=1}^P (1 - z_i z^{-1})$$

(110)

wherein z_i are the roots of LP polynomial $A(z)$ and P is the order of the LP polynomial. In general, the roots of LP polynomial are complex and lie at a radial distance of approximately unity from the origin to the complex z -plane.

20

A new transfer function $\hat{H}(z)$ is determined in

block 114 to attenuate large bandwidth components corresponding to non-vocal tract effects and to emphasize small bandwidth components corresponding to speech.

$\hat{H}(z)$ can be represented in a form parallel to equation 108 by partial fraction expansion as:

$$\hat{H}(z) = \frac{1}{A(z)} - \sum_{i=1}^P \frac{r_i}{(1 - z_i z^{-1})}$$

(112)

wherein residues, r_i , represents the contribution of each component $(1 - z_i z^{-1})$ to the function $\hat{H}(z)$. Residues, r_i , represent the relative gain and phase offset of each component i which can be defined as the spectral tilt of the composite spectrum.

It has been found that spectrum components with large bandwidths correspond to non-vocal tract components and that non-vocal tract components have large residue values.

Normalizing residues, r_i , results in a proportionate contribution of each component i in the spectrum based on its bandwidth. Normalizing residues r_i is performed by setting r_i to a constant, such as unity. For example, if r_i is set to unity the contribution of component i will be approximately:

$$\frac{1}{1 - |z_i|}$$

20

(113)

which is equivalent to the equation:

$$\sim \frac{1}{B_i}$$

5

(114)

10

15

From equation 114 it is shown that the contribution of each component i is inversely proportional to its bandwidth B_i and if component i has a large bandwidth B_i , the value of equation 114 will be smaller than if component i has a small bandwidth B_i . Normalizing of residues r_i can be defined as adaptive component weighting (ACW) which applies a weighting based on bandwidth to the spectrum components of each frame of speech.

Based on the above findings, a new transfer function $\hat{H}(z)$ based on ACW which attenuates the non-vocal tract components while increasing speech components is represented by the following equation:

$$\hat{H}(z) = \sum_{i=1}^P \frac{1}{(1 - z_i z^{-1})}$$

20

(115)

From equation 115, it is shown that $\hat{H}(z)$ is not an all pole transfer function. $\hat{H}(z)$ has a moving average component (MA) of the order $P-1$ which normalizes the contribution of the speech components of the signal.

It is known in the art that cepstral coefficients are used as spectrum information as described in M.R. Schroeder, Direct (nonrecursive) relations between cepstrals and predictor coefficients, Proc. IEEE 29:297-301, April 1981. Cepstral coefficients can be defined by the power sum of the poles normalized to the cepstral index in the following relationship:

$$\ln\left(\frac{1}{A(z)}\right) = \sum_{n=1}^{\infty} c_n z^{-n}$$

15

(116)

wherein c_n is the cepstral coefficient.

Cepstral coefficients, c_n can be expressed in terms of the root of the LP polynomial $A(z)$ defined by equation 106 as:

$$c_n = \frac{1}{n} \sum_{i=1}^P Z_i^n$$

(117)

It is known that prediction coefficients a_i are real. Roots of the LP polynomial $A(z)$ defined by equation 106 will be either real or occur in complex conjugate pairs. Each root of LP polynomial $A(z)$ is associated with the center frequency ω and the bandwidth B_i in the following relationship:

$$Z_i = e^{-B_i + j\omega_i}$$

(118)

center frequency ω_i and bandwidth B_i can be found as:

$$\omega_i = \arctan \frac{\text{Im}(z_i)}{\text{Re}(z_i)}$$

(120)

wherein $\text{Im}(z_i)$ are imaginary roots and $\text{Re}(z_i)$ are real roots and

$$B_i = -\ln|z_i|$$

(122)

Substitution of equation 118 into equation 117 results in cepstral coefficients for speech signal $s(k)$ which can be defined as follows:

$$c_n = \frac{1}{n} \sum_{i=1}^P e^{-B_i n} \cos(\omega_i n)$$

(124)

wherein the n^{th} cepstral c_n coefficient is a non-linear transformation of the MM parameters. Quefrency index n corresponds to time variable k in formula 100 with relative delays ϕ_i set to zero and relative gain G_i set to unity.

From new transfer function $\hat{H}(z)$, a spectral channel and tilt filter $N(z)$ can be determined in block 116. $N(z)$ is a LP polynomial representative of channel and spectral tilt of the speech spectrum which can be defined as:

$$N(z) = 1 + \sum_{i=1}^{P-1} b_i z^{-i}$$

(126)

wherein b represents linear predictive coefficients and P is the order of the polynomial. A FIR filter which normalizes the speech component of the signal can be defined as:

$$\hat{H}(z) = \frac{N(z)}{A(z)}$$

(128)

Factoring LP polynomial $N(z)$ as defined by equation 126 and $A(z)$ as defined by equation 110 results in the new transfer function $\hat{H}(z)$ being defined as follows:

$$\hat{H}(z) = \frac{N(z)}{A(z)} = \frac{\prod_{i=1}^{P-1} (1 - \hat{z}_i z^{-1})}{\prod_{i=1}^P (1 - z_i z^{-1})}$$

(130)

wherein z_i are the roots of the LP polynomial defined by equation 126.

A spectrum with adaptive component weighting (ACW) can be represented by its normalized cepstrum, $\hat{c}(n)$ by the following equation:

$$\hat{c}(n) = \frac{1}{n} \left(\sum_{i=1}^P z_i^n - \sum_{i=1}^{P-1} \hat{z}_i^n \right)$$

(132)

For each frame of digital speech, a normalized cepstrum $\hat{c}(n)$ is computed in block 118. Normalized cepstrum attenuates the non-vocal tract components and increases speech components of a conventional cepstral spectrum. The normalized cepstral spectrum determined from block 118 can be used in classification block 16 or template matching block 18.

Fig. 4 illustrates decomposition of a prior art spectrum of speech for a speaker over a channel from the transfer function $H(z)$. Components labelled 1-4 represent resonances of the vocal tract. A peak in the resonance occurs at a center frequency labeled $\omega_1 - \omega_4$. Each resonance has a respective bandwidth labeled $B_1 - B_4$. Components labeled 5 and 6 represent non-vocal tract effects. Fig. 4 shows that bandwidths labeled B_5 and B_6 representing non-vocal tract effects are much greater than the bandwidths labeled $B_1 - B_4$ for speech components.

Fig. 5 illustrates a decomposition of the spectrum of speech shown in Fig. 4 after application of adaptive component weighting transfer function $\hat{H}(z)$. In Fig. 5, peaks of component 1-4 are emphasized and peaks of components 5 and 6 are attenuated.

Fig. 6A illustrates a prior art spectrum of a speech signal including vocal and non-vocal tract components. Fig. 6B illustrates a spectrum of a speech signal after application of a adaptive component weighting filter. Fig. 6B normalizes peaks 1-4 to a value of approximately 30db for emphasizing the speech components of the signal.

Fig. 7 illustrates the response of moving average filter defined by $N(z)$ for the spectrum shown in Fig. 6B.

Fig. 8 shows a comparison of a spectrum determined from the transfer function $H(z)$ to new transfer function $\hat{H}(z)$. Transfer function $H(z)$ includes a channel effect. Transfer function $\hat{H}(z)$ applies adaptive component weighting to attenuate the channel effect.

A text independent speaker identification example was performed. A subset of a DARPA TIMIT database representing 38 speakers of the same (New England) dialect was used. Each speaker performed ten utterances having an average duration of three seconds per utterance. Five utterances were used for training system 10 in block 16 and five utterances were used for evaluation in block 18. A first set of cepstral features derived from transfer function $H(z)$ were compared to a second set of cepstral features derived from adaptive

component weighting transfer function $\hat{H}(z)$.

Training and testing were performed without channel effects in the speech signal. The first set features of cepstral features from $H(z)$ and the second
5 set of cepstral from $\hat{H}(z)$ had the same recognition rate of 93%.

Training and testing was performed with a speech signal including a channel effect in which the channel is simulated by the transfer function $(1-0.9z^{-1})$.
10 The first set of cepstral features determined from $H(z)$ had a recognition rate of 50.1%. The second set of cepstral features determined from $\hat{H}(z)$ had a recognition rate of 74.7%. An improvement of 24.6% in the recognition rate was found using cepstral features
15 determined by adaptive component weighting.

The present invention has the advantage of improving speaker recognition over a channel or the like by improving the features of a speech signal. Non-vocal tract components of the speech signal are attenuated and
20 vocal tract components are emphasized. The present invention is preferably used for speaker recognition over a telephone system or in noisy environments.

While the invention has been described with reference to the preferred embodiment, this description
25 is not intended to be limiting. It will be appreciated by those of ordinary skill in the art that modifications

may be made without departing from the spirit and scope of the invention.

WE CLAIM:

1. A method for speaker recognition comprising the steps of:

analyzing speech sounds into first spectrum information;

applying weightings to predetermined components from said first spectrum information for producing a normalized second spectrum; and

recognizing said second spectrum by calculating similarity of said second spectrum and a plurality of speech patterns which were produced by a plurality of speaking persons in advance.

2. The method of claim 1 wherein said speech sound is analyzed by dividing said speech sound into a plurality of frames each having a predetermined time period and obtaining LPC coefficients for each of said speech sounds through linear prediction analysis.

3. The method of claim 2 wherein a first weighting is applied to a first predetermined component and a second weighting is applied to a second predetermined component.

4. The method of claim 3 wherein said first weighting increases said first component and said second





weighting attenuates said second component.

5. The method of claim 4 wherein said first and second weightings are determined for each of said frames.

6. The method of claim 5 wherein said second component has a large bandwidth.

7. The method of claim 6 wherein said first and second weightings are determined by the steps of:
determining roots of said LPC coefficients,
each said root including a residue component; and
normalizing said residue component of each of said roots.

8. The method of claim 7 wherein said second spectrum is defined by the transform:

$$\hat{H}(z) = \sum_{i=1}^P \frac{1}{(1 - z_i z^{-1})}$$

wherein P is the order of analysis, and z_i is a complex root representing the bandwidth and frequency of said speech sound.

9. The method of claim 8 further comprising





the step of:

determining a normalized cepstrum for said second spectrum of speech.

10. The method of claim 9 wherein said cepstrum is defined by:

$$c(n) = \frac{1}{n} \left(\sum_{i=1}^P z_i^n - \sum_{i=1}^{P-1} \hat{z}_i^n \right)$$

wherein P is the number of coefficients z_i is a complex root of said second spectrum.

11. The method of claim 10 further comprising the step of:

storing said second spectrum in a classification means as said plurality of speech patterns.

12. The method of said claim 11 further comprising the step of:

determining said similarity of said second spectrum with said speech patterns by matching said second spectrum with said stored second spectrum in said classification means.

13. A system for speaker recognition

This Sheet Has Been Replaced





comprising:

means for converting a speech signal into a plurality of frames of digital speech;

speech parameter extracting means for converting said digital speech into a series of first spectrum parameters;

speech parameter enhancing means for applying adaptive weightings to said first spectrum parameters for producing a normalized second spectrum; and

evaluation means for determining a similarity of said second spectrum with a plurality of speech samples which were produced by said speaking persons in advance.

14. The system of claim 13 wherein said plurality of speech samples are produced by said second spectrum.

15. The system of claim 14 wherein said first spectrum is produced by filtering said speech signal through an all pole LPC filter.

16. The system of claim 15, wherein said speech signal includes vocal and non-vocal tract components and wherein said speech parameters enhancing means includes a moving average filtering means for

This Sheet Has Been Replaced





attenuating said non-vocal tract components and for increasing said vocal tract components.

17. The system of claim 16 wherein said system further comprises:

vector quantization means for quantizing said plurality of speech samples; and

means for storing said quantized samples in a code book.

18. The system of claim 17 wherein said evaluation means comprises template matching means for determining the similarity of said second spectrum with said plurality of speech samples by matching said second spectrum with said samples stored in said code book.

19. The system of claim 18 wherein said second spectrum is a normalized cepstrum $\hat{c}(n)$ which is defined by

$$\hat{c}(n) = \frac{1}{n} \left(\sum_{i=1}^P z_i^n - \sum_{i=1}^{P-1} \hat{z}_i^n \right).$$

This Sheet Has Been Replaced

