Very Low Rate Speech Coding Algorithm Based on Biomimetic Pattern Recognition

Yanyan Yin\textsuperscript{a,b}, Ye Yin\textsuperscript{a,*}, Xiangnong Wu\textsuperscript{a}, Wei Hu\textsuperscript{a}

\textsuperscript{a}College of Information, Mechanical and Electrical Engineering, Shanghai Normal University
Shanghai 200234, China
\textsuperscript{b}Shanghai Zenkore Electronic Technology Co., Ltd., Shanghai 201100, China

Abstract

Based on the speech coding technology, in this paper we propose a new method encoding the speech signal by the bionic pattern recognition technique. Using this new encoding method, the text message can be received after using biomimetic pattern recognition of the speaker’s voice. And the information related to the speaker’s individual characteristics can be obtained by “comparison” between the speaker’s voice and the standard voice which corresponding to the text message. Then the text message and individual characteristics information are encoded before transmission. Compared with the application of traditional speech recognition in low-rate speech coding, the scheme using biomimetic pattern recognition is single template identification with a faster recognition speed and lower instructions (keywords) false acceptance rate, which nearly reaches the limits of the voice coding rate. This new method is anticipated to have great potential application in military communications, underwater communications, secure communications and other demands under special conditions.

Keywords: Speech Recognition; Coding Rate; Very Low Bit Rate Speech Vocoders; Biomimetic Pattern Recognition

1 Introduction

According to the coding rate, speech coding can be divided into five categories: high-speed 32 Kb/s or more; medium high rate from 16 to 32 Kb/s; low rate of 1.2-4.8 Kb/s; and very low speed speech coding with bit rate lower than 1.2 Kb/s [1]. The so-called voice compression coding means its coding rate being 16 Kb/s or less. Low-rate speech coding is an important direction in modern research and development of speech coding technology. The algorithms used in the existing low rate speech coding system show a wide range of cross-penetration tendency.

*Project supported by Project of Scientific Research Programs of Science and Technology Commission of Shanghai (No. 10dz1110900), and Project of International Technology Transfer of Science and Technology Commission of Shanghai (No. 10510708400).
*Corresponding author.
Email address: yinye@188.com (Ye Yin).

1548–7741 / Copyright © 2013 Binary Information Press
January 1, 2013
There are mainly four models [4], such as Mixed Excitation Linear Prediction (MELP), Multi-band Excitation (MBE), Waveform Interpolation (WI) and Sinusoidal Transform Coding (STC). Low rate speech coding has broad application prospects in confidential communication, voice mail, network communications, IP telephony, etc.

2 Very Low Speed Speech Coding

2.1 Overview [3]

In general, there are three major research directions for current speech coding technology. The first is the speech signal coding without compression. Currently, the transmission bandwidth is growing rapidly due to the development in fiber access systems. This greatly decreases the cost of information transmission system. With the increasing demands for high-quality voice, and compared with the transmission saved cost, the expenditure in speech coding becomes uneconomical. The cost for speech systems without compression transmission becomes comparably much lower. The second direction lies in the variable rate speech coding. The development of the third generation mobile communication makes much higher requirements for voice coding algorithms. It not only requires a low coding rate to increase the system capacity, but also requires a high quality synthesis to ensure the call quality. Under this situation, variable rate speech coding methods has been proposed, which can dynamically adjust the encoding rate by obtaining a flexible compromise between the synthesis of voice quality and system capacity to maximize the system performance. The third direction is much lower-rate speech coding technology. The modern communications systems want higher utilization efficiencies and less channels. Hence, in this paper, we mainly study the very low rate speech coding technology, which is also an indispensable research direction of voice communications.

With the rapid development and progresses in a lot of fields, such as computer technology, microelectronics, signal processing, phonetics, semantics, neurophysiology, and coding theories, many breakthroughs have being made in speech coding technology. Along with this, varieties of practical speech coding standards are established [5]. The various international speech coding standards also reflect the development level of the speech coding technology. These standards are based on many major performance indicators, such as practical application backgrounds, quality of coding, coding rate, coding delay, the complexity and robustness of algorithms, and so on. So we can get an over-all balance and obtain an optimal choice. In Fig. 1, the relationship between encoding quality and encoding speed is quantitatively plotted.

2.2 Theory Basis [6]

From the viewpoint of theoretical analyses of acoustic theories, there exists a big space for speech signal compression. There are two major reasons.

One reason is that, the speech signal contains a large amount of redundant information. It can be shown in the following four aspects. (1) Voice signal samples have strong correlations between these samples, where short-term spectrum is not flat. (2) The voiced speech segment has a quasi-periodic feature. (3) The shape and the change rate of the channel are limited. (4) The value of the transmitted code is of non-uniform probability distribution.
Another reason lies in that, we can take advantages of the psychological characteristics of the human ears. (1) For different frequencies of sound, human ears have different sensitivities. Normally we are more sensitive to low frequency than to high frequency. (2) The phase of speech signals has no influence to our human ears. (3) The human ears have “masking effect”. In essence, the voice compression coding is managing to remove the redundant information based on the psychological characteristics of human ears. Therefore, it can reach the purpose of voice compression.

### 2.3 Encoding Rate Limit

From the viewpoint of information theory, 21 consonants and 38 vowels of Chinese composed of about 417 syllables. With four tones, there are 1600 tone syllables. But some syllables do not correspond to any character. According to statistic data, there are 412 Chinese words almost no tone-syllable, words with one tone-syllable are 1282 [2]. If the syllable rate is 2-5 syllables/sec, its information entropy can be shown in Equation (1).

\[
I = \log_2(412 + 1282)^5 \approx 50(b/s)
\] (1)

We can see that in a writing text that is equivalent of voice speaking under normal speed, the information entropy is roughly 50 bit/s. Actually, the bottom line of voice information is much higher than 50 bit/s. It is because that the above estimation does not take into account the speaker’s personality, mood and voice strength and other information.

From the text viewpoint, each Chinese character can be described by 16 bits. And the average speed of standard announcer is 4 words per second. Hence, the average information rate is about 64 bits per second. Differently, seven bits is enough to express a letter in English, and the voice
communication rate is 125 English words per minute. Assuming an average of seven letters form one word, the information entropy can be about 100 b/s.

So the limiting rate of voice compression coding can be deemed as within the range of 50-100 bit/s. Now only language content can be transmitted, pitch and other important information has been lost. There is a considerable span (about 640 times) between the limiting encoding rate and the standard encoding rate (64 kb/s), which is of great significance for speech coding’s research and practice [6]. We can estimate the encoding rate of the Chinese identification synthesizer vocoder according to the results of existing research. Among all the information that needs to be transferred, pitch contour information coding is a very important message. Using scalar quantization, each frame (10 ms) with 8bit encoding of pitch, the coding rate will be 800 b/s. Here we use the vector quantization technology to reduce the bit rate to 8bits per syllable, which are equivalent to 16-40 bits per second (or 2-5 syllable pronunciation per second), and 11bits are needed for the alphabet and tone of each syllable coding, every syllable duration and the energy factor need 5 bits, and the sound league sign needs 1 bit per second. So it requires a total transmission of 30bits per syllable. And the bit rate is about 60-150 bit/s [1].

3 Speech Coding Algorithm Based on Biomimetic Pattern Recognition

Based on information theory, the lower limit of speech coding information rate is 50 bit/s. Many studies show that a various of current low rate speech coding algorithms used in the analyses and LP-based syntheses are difficult to meet the requirements for the lower bit rate down to 400 b/s. And the voice quality provided cannot reach an acceptable level. The reason is that such analyses and syntheses of type vocoder coding unit of voice signal is only one or several frames, which are characterized by endless variations as each frame time about 10-30ms. But these vocoder have only one the length of finite symbol set to encode, which is bound to make the reconstruction of voice signals have a considerable distortion, resulting in unacceptable quality. Therefore, we can only adopt the speech recognition [8, 12, 13, 14] and synthesis technology to create synthetic recognition vocoder to reduce the speech coding rate down below 400 b/s, which closing to the lower bound of 50 bit/s.

3.1 System Overview [1]

Fig. 2 shows the basic principles of the identification and synthesis vocoder, which using biomimetic pattern recognition for speech recognition and synthesis technology to encode speech primitives. Voice of the primitives can be phonemes, syllables or words, any phoneme or syllable of a language is a finite set of numbers, which can be used as a primitive encoding, achieving unlimited vocabulary speech coding. Biomimetic pattern recognition for voice recognition technology is used of these vocoder in the transmitter to identify and primitive encoding. And then the receiver re-synthesized speech according to the voice primitive code (text messages) and some additional rhythm information (semantic information) it received. Because there are very few transmission parameters need to be coded, its transmission rate can be very low, and the quality of the synthesized voice recovered at the receiving end will be quite good.
3.2 Algorithm \[6, 11\]

In feature space $\mathbb{R}^n$ of Biomimetic Pattern Recognition, any type of things (such as the $A$ class), all the set of points formed by the “shape” mapped in $\mathbb{R}^n$ (must be a continuous map) is considered as a closed set, called $A$ collection. According to the specific application object of Biomimetic Pattern Recognition, $A$ can be different dimensions of “manifold”. The identification process is to determine whether the “shape” in feature space $\mathbb{R}^n$ mapped by “recognized things” is a part of the set $A$. The $n$-dimensional space geometry must be structured in a feature space $\mathbb{R}^n$ that can cover $A$. Therefore, the essence of the algorithm is to make a proper super-ellipsoid in the feature space.

In the two-dimensional space, the super-ellipsoid is an oval. To ensure the equation of the ellipse, the values of the ellipse parameters $a$, $b$ and $\theta$ should be known, where $a$, $b$ are the two semiaxis of the ellipse, and $\theta$ is the rotation angle. And the crucial lies in how to determine the length of the two semiaxis.

In the first step, we fit curve using least square method. Then the coefficients of polynomial expression can be obtained. According to the sample set, we first use one square line fitting. It is assumed that there are $m$ samples of two-dimensional space, they can be represent by $(x_1, y_1)$, $(x_2, y_2) \cdots (x_m, y_m)$ in the plane XOY. We suppose that $y$ is the line function of $x$.

$$y = kx + d$$ \hfill (2)

where $k$, $d$ are undetermined coefficients. Generally speaking, this $m$ points cannot be located in the same line. The following equation reflects the error between the line we got above and the
true value.

\[ \varepsilon_i = y_i - (kx_i + d) \quad i \in 1, 2, 3, \ldots, m \]

According to the minimum mean square error criterion, it should be made get its minimum.

\[ E(k, d) = \sum_{i=1}^{m} \varepsilon_i^2 = \sum_{i=1}^{m} (y_i - kx_i - d)^2 \]  

(3)

Based on the extreme value principle, we have:

\[ \frac{\partial E}{\partial k} = -2 \sum_{i=1}^{m} x_i (y_i - kx_i - d) = 0 \]

Solving simultaneous equations, we have

\[ k = \frac{m \sum_{i=1}^{m} x_i y_i - \sum_{i=1}^{m} x_i \sum_{i=1}^{m} y_i}{m \sum_{i=1}^{m} x_i^2 - \left( \sum_{i=1}^{m} x_i \right)^2} \]  

(4)

\[ d = \frac{\sum_{i=1}^{m} \sum_{i=1}^{m} y_i - \sum_{i=1}^{m} x_i \sum_{i=1}^{m} x_i y_i}{m \sum_{i=1}^{m} x_i^2 - \left( \sum_{i=1}^{m} x_i \right)^2} \]  

(5)

In the second step, after getting the slope and excursion values of the line, we can determine the direction vector of the axis of the oval respectively: a-axis \((1, k)\), b-axis \((k, -1)\). Then we can find all the two-axis direction cosine:

\[ \cos \alpha_1 = \frac{1}{\sqrt{(1 + k^2)}}, \quad \cos \beta_1 = \frac{k}{\sqrt{(1 + k^2)}} \]

\[ \cos \alpha_2 = \frac{k}{\sqrt{(1 + k^2)}}, \quad \cos \beta_2 = \frac{-1}{\sqrt{(1 + k^2)}} \]

where \(\alpha_1, \beta_1\) are the direction angles of a-axis in the OXY coordinate system, \(\alpha_2\) and \(\beta_2\) are the direction angles of b-axis in the OXY coordinate system. On the straight line Equation (2), we find all the projection points of the sample points \(\sum (x'_i, y'_i)\). From these projection points, we can get the two points farthest apart \((x'_i, y'_i)\) and \((x'_j, y'_j)\). The distance between them is \(d_{\text{max}}\), so \(a = d_{\text{max}}/2\).

And the center coordinates of the ellipse is

\[ (x_0, y_0) = \left( \frac{x'_i + x'_j}{2}, \frac{y'_i + y'_j}{2} \right) \]

Then we calculate the distance from each point to the fitted line. The maximum distance of which is \(b\) will be

\[ b = \max \left( \frac{|kx_i - y_i + d|}{\sqrt{k^2 + d^2}} \right) \]
In the third step, all parameters of the oval have been defined. Thus the elliptical equation can be expressed as

\[
\frac{x'^2}{a^2} + \frac{y'^2}{b^2} = 1
\]  

(6)

where,

\[
\begin{bmatrix}
  x' \\
  y'
\end{bmatrix} = \begin{bmatrix}
  \cos \alpha_1 & \cos \alpha_2 \\
  \cos \beta_1 & \cos \beta_2
\end{bmatrix} \begin{bmatrix}
  x \\
  y
\end{bmatrix}
\]  

(7)

The discriminant function for identification is

\[
\Phi(x', y') = 1 - \left( \frac{x'^2}{a^2} + \frac{y'^2}{b^2} \right)
\]  

(8)

When \( \Phi > 0 \) means that the input \((x', y')\) is covered into the area of the ellipse.

After using biomimetic pattern recognition method to divide, in Fig. 3, we can see the four types of samples’ distribution in two-dimensional feature space, where, different regions in the elliptical points is from different types of samples.

For the three-dimensional space, the algorithm in which the goal is to make an ellipsoid. The key is to determine its length and direction vector of the three-axis. Modeled on the one-dimensional linear fitting method, assuming that there is a plane

\[
z = k_1 x + k_2 y + k_3
\]  

(9)

In accordance with minimum mean square error criterion, we have

\[
E = \sum_{i=1}^{m} (z_i - k_1 x_i - k_2 y_i - k_3)^2
\]  

(10)
\[ \frac{\partial E}{\partial k_1} = \frac{\partial E}{\partial k_2} = \frac{\partial E}{\partial k_3} = 0 \] (11)

Solving equations, we get \( k_1, k_2, k_3 \) and the normal vector of the plane is \( L = k_1 k_2, -1 \). \( L \) is also a certain vector of ellipse axis, which we call a-axis. Take the maximum value of all the sample points to the plane as the semiaxis length, making all the sample points project onto the defined plane, then in this plane an ellipse can be determined, which method is exactly the same to the ellipse defined in two-dimensional space. This can determine the ellipsoid axes in the direction of the three vectors.

Assuming three axes of \( a, b, c \), corresponding to the direction cosine \( \cos \alpha_1, \cos \beta_1, \cos \gamma_1; \cos \alpha_2, \cos \beta_2, \cos \gamma_2; \cos \alpha_3, \cos \beta_3, \cos \gamma_3 \) respectively. The equation of the ellipsoid can be expressed as Equation (12).

\[ \frac{x'^2}{a^2} + \frac{y'^2}{b^2} + \frac{z'^2}{c^2} = 1 \] (12)

where,

\[
\begin{bmatrix}
  x' \\
  y' \\
  z'
\end{bmatrix} =
\begin{bmatrix}
  \cos \alpha_1 & \cos \alpha_2 & \cos \alpha_3 \\
  \cos \beta_1 & \cos \beta_2 & \cos \beta_3 \\
  \cos \gamma_1 & \cos \gamma_2 & \cos \gamma_3
\end{bmatrix}
\begin{bmatrix}
  x - x_0 \\
  y - y_0 \\
  z - z_0
\end{bmatrix}
\] (13)

Similarly, the discriminant function is Equation (14).

\[ \Phi(x, y, z) = 1 - \left( \frac{x'^2}{a^2} + \frac{y'^2}{b^2} + \frac{z'^2}{c^2} \right) \] (14)

If it is \( n \)-dimensional space, then the point is to determine \( n \) axes of an \( n \)-dimensional super-ellipsoid. Its method is the same as in three-dimensional space. First, finding a \((n - 1)\) dimensional hyper-plane according to the minimum mean square error criterion, whose normal vector is the direction vector of an axis. Then all the sample points are projected on this hyper-plane, the projection point can be regarded as \((n - 1)\) dimensional space of sample points, which achieves the \( n \)-dimensional space to \((n - 1)\) dimensional space of dimension reduction. Repeating this projection process, we decrease the \( n \)-dimensional space to the three-dimensional space. Finally, we can write \( n \)-dimensional super-ellipsoid equation and discriminant function as the following Equation (15).

\[ \frac{e_1^2}{r_1^2} + \frac{e_2^2}{r_2^2} + \ldots + \frac{e_n^2}{r_n^2} = 1 \] (15)

where,

\[
\begin{bmatrix}
  e_1' \\
  e_2' \\
  \vdots \\
  e'_n
\end{bmatrix} =
\begin{bmatrix}
  \cos \theta_1 & \cos \theta_2 & \ldots & \cos \theta_n \\
  \cos \theta_1 & \cos \theta_2 & \ldots & \cos \theta_n \\
  \vdots & \vdots & \ldots & \vdots \\
  \cos \theta_1 & \cos \theta_2 & \ldots & \cos \theta_n
\end{bmatrix}
\begin{bmatrix}
  e_1 - e_1^0 \\
  e_2 - e_2^0 \\
  \vdots \\
  e_n - e_n^0
\end{bmatrix}
\] (16)

\[ \Phi = 1 - \left( \frac{e_1^2}{r_1^2} + \frac{e_2^2}{r_2^2} + \ldots + \frac{e_n^2}{r_n^2} \right) \] (17)

In summary, the biggest advantage of the biomimetic pattern recognition applied to the key words identification is that the vocabulary sheet can be automatically rejected. In another words,
biomimetic pattern recognition’s active can reject foreign words. This reduces the error rate. At the same time, this rejection to external vocabulary knowledge does not increase the leakage rate. This characteristic of biomimetic pattern recognition fits well with the evaluation criteria of low-rate speech coding identification system.

4 Summary

The main idea of this paper is: the text message will be received after using biomimetic pattern recognition of the speaker’s voice, and the information about the speaker’s individual characteristics can be obtained by “comparison” between the speaker’s voice and the standard voice which corresponding to the text message, then encode the text message and individual characteristics information before transmission. The main work of this article focuses on the voice coding technology. Given the limits of low rate speech coding, based on this application we give the Biomimetic pattern recognition algorithm from the mathematical point of view. Based on its faster recognition speed and lower instructions (keywords) false acceptance rate, and other merits, experiments also show that the use of biomimetic pattern recognition in the low rate speech coding systems is a valuable attempt. Due to limited space, the process of personal voice characteristic information extraction and speech synthesis of the receiver will be described in subsequent papers.

References

[7] Li Guo, Lihong Guo, Bingxi Wang, New developments and applications of very low bit rate speech coding, Voice Technology, 06, 2005, 38-41