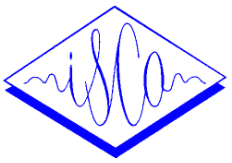

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

A NEW TONE CONVERSION METHOD FOR MANDARIN BY AN ADAPTIVE LINEAR PREDICTION ANALYSIS

Wang Congxiu Li Qihu Zhao Guoying Yin Li Hao Shuai Meng Da

Institute of Acoustics, Academia Sinica, Beijing, China (100080)

ABSTRACT

Conventional frame-based linear prediction coding(LPC) does not always de-convovle the speech signal into the vocal tract response and the voice source excitation signal exactly. In order to get better results, an adaptive linear prediction coding analysis(ALPC) method is proposed in this paper. The vocal tract responses obtained from ALPC analysis are used to synthesize a vowel syllable that has a different tone. During the synthesizing procedure, the voice source excitations are superseded by the improved LF-4 model, and the pitch-synchronized synthesizing method is used. The hearing experiments indicate that the synthesized speech by ALPC analysis, which has a different tone from the original one, has a high intelligibility and a better voice quality than that by the traditional LPC analysis method.

1. INTRODUCTION

Mandarin is a language that has four tones, which have the function to distinguish different meanings. The intelligibility tests of mandarin show that tone is the most susceptible factor to be perceived and its ability to anti-disturbance is also very strong. In the non-interactive models of speech production, the tone due to the vibration of the vocal cord and the formants determined by the vocal tract are two independent factors. The two factors can be derived by de-convolution method, such as LPC, cepstrum, etc. Different tones are due to different pitch trajectory, namely different pitch contour. Among the four tones of the same vowel, the maximum variation of the pitch can amount to 5/3 octave, but the locations of the formants remain almost the same (Zhang Jialu, 1979). So the changing mode of pitch is a very important factor that determine the tone and naturalness of the synthesized speech. By changing the pitch contour of the synthesized speech, and at the same time, keeping the impulse responses intact, we can converse one tone into another tone of the same syllable.

2. ADAPTIVE LINEAR PREDICTION CODING ANALYSIS METHOD

As we all know, the LPC method is a very useful de-convolution method by which we separate the speech signal into the vocal tract response and the voice source excitation. The traditional LPC analysis method still has some drawbacks, and the results, especially the vocal tract response, are largely affected by the following factors: (1) the position of the analysis frame, and (2) the length of the analysis window (H. Fujisaki, 1987, D. G. Childers, 1991). This is mainly because that the speech signal is non-stationary, and the shape of the vocal tract may change slowly during the same syllable. On the other hand, our experiment shows that the accuracy of estimated parameters is also sensitive to the order of the LPC analysis. The inverse filter A(z), which represents the vocal tract response, sometimes is not a minimum phase system if different LPC orders are utilized on the window with the same length, just as Fig. 1 shows. Here the impulse response is obtained from the following

equalizations. And the procedure is similar to that seeking the coefficients of the LPC cepstrum.

$$A(z)=\frac{1}{1-\sum_{k=1}^{P}a_k z^{-k}} \qquad (1)$$

An all-pole filter has an infinite impulse response, and here we use h(n) to represent the time-domain impulse response of the vocal tract. We get,

$$A(z)=\sum_{n=1}^{\infty}h(n)z^{-n} \qquad (2)$$

Thus, we have

$$\frac{1}{1-\sum_{k=1}^{P}a_k z^{-k}}=\sum_{n=1}^{+\infty}h(n)z^{-n} \qquad (3)$$

By comparing the two sides of equation (3), we can get following relation between the time-domain impulse response h(n) and the coefficients of the linear prediction filter a_k.

$$\begin{cases} h(1)=1 & (n=1) \\ h(n)=-\sum_{k=2}^{n}a(k)h(n-k+1) & (1<n\le P+1) \\ h(n)=-\sum_{k=2}^{P+1}a(k)h(n-k+1) & (n>P+1) \end{cases} \qquad (4)$$

Thus, from equation (4), we can get the time-domain impulse response. Fig. 1 shows the impulse responses of different LPC orders of the same window from the same start-point of vowel /a:/, and the start-point is an arbitrary one. The syllable, which has a falling-rising tone, is pronounced by a young female. The sampling rate is 11.025 kHz, the Hamming window is used, and the window length is 300. The signal is pre-emphasized before LPC analysis, and the LPC order is 12, 14, 16, and 18 respectively.

From Fig 1, we can see that among the four different LPC orders we chose, the impulse response

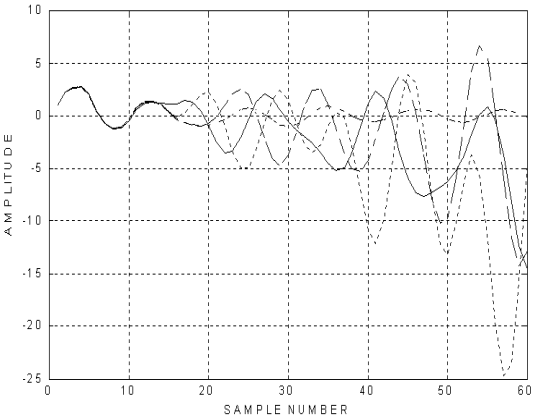

Fig 1 The impulse responses of different LPC orders of the same frame
(solid: R=12, dotted: R=14, dashdot: R=16, dashed: R=18, R is the order of LPC analysis)

of the vocal tract converged only when the LPC order is equal to 16. And so we can conclude that the results of LPC analysis are sensitive to the LPC orders. Unfortunately, the LPC order that offer desirable results does not remain the same. It changes among the same syllable, and with no obvious rule.

Just as the above shows, sometimes, the impulse response does not decrease to zero with time on, and this does not fit to actual situations. Even a single of this kind of impulse response will deteriorate the quality of the synthesized speech when it is used to synthesize speech. Our synthesis experiments show that if the impulse response decrease quickly, the convolution of the prediction residual signal and the impulse response will fit the original signal with little difference. So we can conclude that if the time-domain impulse response decrease quickly, the vocal tract response and the voice source excitation can be said to have been de-convolved well.

Based on the above analysis, we proposed an adaptive linear prediction method. Here, ‘adaptive’ means that both the frame length and the LPC order can be changed during the analysis process, until an ideal result is obtained.

In order to describe the decreasing rate of the

impulse response of the vocal tract, we give the following definition, the tail-head-ratio (THR).

$$THR = \frac{\sum_{i=0}^{L-1} abs(h(N-i))}{\sum_{i=1}^{L} abs(h(i))} \qquad (5)$$

In equation (5), h(n) is the impulse response of the vocal tract. It is cut from the first one to the length of N. If the THR of a speech frame is less than a given threshold THR_0, we can think that the de-convolution results meet our needs.

The following Fig. 2 is the flow chart of the ALPC analysis. Before this process, the S/U/V segmentation of the speech signal is made, and only the voiced part is processed. The analysis need not to be pitch-synchronized, for the parameters that represent the shape of vocal tract change slowly during the same syllable.

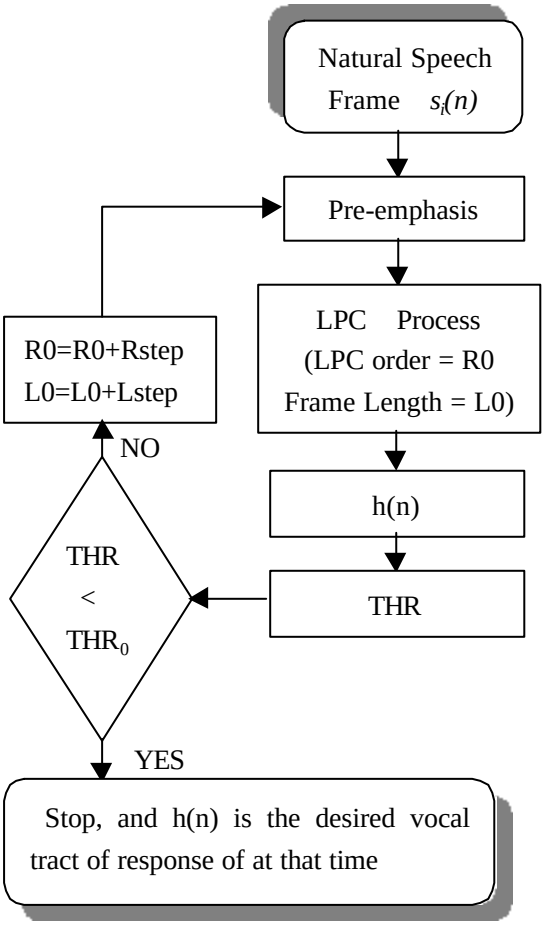

Fig. 2. The flow chart of ALPC analysis

3. TONE CONVERSION

3.1 Voice Source

An accurate voice source excitation model is essential if we want to synthesize syllables with high quality. Much of our emphasis is put on seeking a qualified voice source. There are several good voice excitation models, such as the LF-4 derivative voice excitation model and its improved forms (R.Carlson,1989, G. Fant, 1993, D. H. Klatt, 1990), and the 6th polynomial model used by Milenkovic to approach the glottal waveform (P. H. Milenkovic, 1993). We choose the LF-4 voice source model as the excitation signal to supersede the prediction residual, because it has a set of parameters with definite physical meaning, and the parameters can be easily controlled. A little difference from the original model, we use the time domain parameters, namely t_p, t_e, t_a and t_c, directly. These four parameters can be automatically estimated from the inverse filtered speech signal(Helmer Strik, 1998). In order to simplify the computation, the negative maximum value, E_e, is also used as a parameter. The parameters of the LF-4 model is not only related to different speaker types, such as adult male, adult female, and children(R. Carlson,1989), but also related to different voice types, such as modal voice, vocal fry voice and breathy voice(D. G. Childers, 1995). And to some extent, these voice source parameters also represent the ‘ individuality’ of the speaker.

Besides the LF-4 model, we add some high frequency noise with low amplitude, for the speech signal is hard to be divided into the generally used two groups: voiced and unvoiced. Even the absolute voiced frame with a definite period has non-periodical characteristics in its high frequency end. The speech coding research, such as MBE, also proves this.

3.2 Tone Model

There are totally four basic types of tone in Mandarin, the high and level tone, the rising tone,

the falling-rising tone and the falling tone. The benchmark value of tone varies from person to person. But for the same person, the benchmark of tone is about the same. Based on this principle, we can draw out a set of values that represent a given tone, and that set of value can be pre-stored in the computer.

3.3 Amplitude Envelope

The amplitude envelope is defined as equation (6). Different tones have different amplitude envelope, and the amplitude envelope shows the energy variation in the duration of a syllable. Their common tendency is that the envelope goes up first and then goes down, but they discriminate with each other in details.

$$AE_j = \frac{1}{P_j}\sum_{j=1}^{P_j} abs(s_j(n)) \qquad (6)$$

In equation (6), AE_j is the mean amplitude envelope of the jth period, P_j is the sample number in the jth period, $s_j(n)$ is the samples of the jth period.

3.4 Synthesis

The impulse response of the vocal tract and the LF-4 derivative voice source excitation is used to synthesize the syllable with the target tone. The synthesis is pitch-synchronized. In order to smooth the border of two sequential pitch period, an exponentially decaying window is used to concatenate one pitch period with the following one. The length of the synthesized speech is about the same as the original one, for there is no statistical difference in time length among the four tones of a given syllable.

The hearing experiments show that the synthesized speech, with a totally different tone, has a high voice quality and intelligibility.

4. SUMMARY

In this paper, we proposed a tone conversion method for Mandarin. By an adaptive LPC analysis method, we can get a relatively accurate impulse response of the voiced speech signal. The improved LF-4 derivative voice source model is used in synthesis to replace the excitation residual signal. By suitable control over the pitch contour, amplitude envelope, and the time length, we can converse the speech signal into a totally different tone. The method will also do good to speech coding, speech normalization, and those with speech disorders.

REFERENCE

1. D. G. Childers, and Chieteuk Ahn,Modeling the glottal volume-velocity waveform for three voice types, J. Acoust. Soc. Am. 97(1), 505-519,1995.

2. D. G. Childers, and Ke Wu, Gender Recognition from speech. Part 2: Fine analysis, J. Acoust. Soc. Am.,90(4), 1841-1856, 1991.

3. Helmer Strik Automatic parametrization of differentialted glottal flow: Comparing methods by means of synthetic flow pulses, J. Acoust. Soc. Am.103(5), 2659-2669,1998.

4. Zhang Jialu, Qi Shiqian and Lü Shinan, A semi-dynamic method for spectrographic analysis of vowels. J. Acous. Soc. China, 1(1979),23-29.

5. P. H. Milenkovic, Voice source model for continuous control of pitch period, J. Acoust. Soc. Am, 93(2),1993, 1087-1096.

6. R. Carlson, G. Fant, C. Gobl, Voice source rules for text-to-speech synthesis, Proc. ICASSP-89, 223-226,1989.

7. D. H. Klatt, and L. C. Klatt, Analysis, synthesis, and perception of voice quality variations among female and male talkers, J. Acoust. Soc. Am. 87(2), 820-856,1990.

8. H. Fujisaki, M. Ljungqvist, Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform, Proc. IEEE ICASSP-87,637-640,1987.

9. G. Fant Some problems in voice source analysis, Speech Communication, 13(1993),7-22,1993.