The study and development of the Speech Labeling Tool

Tao Jianhua Cai Lianhong

Information group, Department of Computer Science and Technology
Tsinghua University, Beijing, 100084
E-mail: tjh@tts.cs.tsinghua.edu.cn clh-dcs@mail.tsinghua.edu.cn
Phone:8610-62782406, Fax:8610-62772001

ABSTRACT

In the paper, an automatic speech signal labeling tool, which can be run in Windows95/98/NT system, is described. The current version of tool provides a wide range of functionality, which can be divided into four main parts: wave management, speech analysis, prosodic labeling and speech processing. Especially, the tool is very suitable for the analysis and prosodic labeling of the speech database. Moreover several new methods are adopted to improve the accuracy in speech parameter calculation. The tool is a powerful, extensible, object-oriented system, allowing researchers to rapidly build and configure customized speech analysis functions, the results can also be shared by users. We now use the tool widely in almost all aspects of our speech analysis, speech database labeling and synthesis research.

KEYWORDS: Labeling tool, Speech analysis, Prosodic labeling, Speech processing

1. INTRODUCTION

Along with the development of the speech processing technology, speech synthesis systems have been used more and more widely in our lives. But, there are still some problems in them, such as the naturalness of the sound generated by the synthesis system still need to be improved. Recently, many companies or academies have established large databases, which will be helpful to greatly improve the quality of the synthesis system.

The database usually includes the speech data and the corresponding text. Since it is such a teadious job in database labeling, a powerful tool for automatic analysis and labeling of the speech database is particularly necessary. There are several existing tools great towards the development of human language technologies. But for the research of Chinese speech synthesis, they usually cannot give us enough detailed information. In this paper, a labeling tool, which has powerful functions and is very suitable for the analysis and labeling of segmental and prosodic information of Chinese speech database, is described. Moreover, the tool is designed as an opened platform, which means the functions, developed by users, can be integrated into the tool expediently, catering for different needs.

In database design, we usually find that the speech data is very large. To manage them efficiently, we add wave management functions in the tool. Through these functions, the tool can read, save, record, play and edit the wave data, and also can compress and decompress the speech with ADPCM coding. The applications indicate these functions are much useful in speech data management.

Some acoustic parameters of speech, such as wide-band spectrogram, formants, energy and zero crossing etc., are much important for researchers to apply the speech database. The parameters are very useful for us to get the detailed information of the speech, especially in studying features of the speech. In section 3, we describe the speech analysis functions and the calculation method of them.

Since the prosodic features do the most important role in the study and the developing the speech synthesis system, they deeply influence the natureness of the synthesis results. The quality of the prosodic labeling of the speech will restrict the applications of the database. Thus, prosodic labeling is the kernel job of the tool. In section 4, we will detail the prosodic labeling method and corresponding functions.

In order to help us study the prosodic rule for the speech synthesis system and to optimize the speech database, the tool is integrated some other useful functions, such as modifying the magnitude and pitch of the speech, re-synthesizing the speech and filtering speech with Chebychev or Butterworth filter, etc.. They will be detailed in section 5. Furthermore, some successful applications of the tool will be introduced in section 6.

2. WAVE MANAGEMENT

Managing and editing the speech data is the major task

This research is supported by National 863 High Technology Project and National Natural Science Foundation of China (69875008)

in wave management functions. In the current version, the tool is able to identify several speech file formats, such as 'WAV', 'VOC', 'ADPCM' etc., and also can process the speech with two channels. Wave data will be shown in the window, and can be arbitrarily zoomed out or zoomed in time scale. Through some operations, such as cutting, copying and pasting, etc., the wave can be edited by the mouse or keyboard. Thus, you are able to combine the wave in your own options. It will be much useful for us to optimize the speech database.
Moreover, the tool also supplies us recording and playing functions, which can be thought to be the basic parts of the tool.

3. SPEECH ANALYSIS

The requirements from users are in a large range, we cannot confirm which acoustic parameters they might require. Thus, in addition to prosodic labeling of the database, the tool must be able to supply several acoustic parameters from the speech. Here, spectrum and zero crossing rate of the speech are analyzed and displayed in the tool. If it is needed, the LPC parameters and cepstrum parameters of the speech can also be got.
In speech analysis process, all waves are framed by the Hanning window with the length of 200ms. Usually, spectrum takes on two aspects, wide-band spectrum and narrow-band spectrum (also called formant spectrum). Figure 1 shows the result of the wide-band spectrogram calculated from the sentence, 'yi1jiu3jiu3er3nian2, ti3yu4jian4zhu4jiang3'. In the wide-band spectrogram, we can see that both the resonant frequency and the harmonic frequency are all displayed clearly.

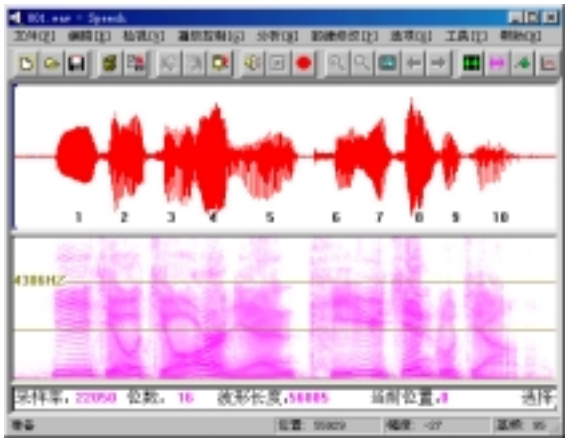

Fig1. Wide-band spectrogram of speech

But the formants of the speech directly reflect the vocal tract of the speaker and they are much important in the study of speech synthesis and speech recognition. The calculation of the formants is not a easy job. In the tool, the whole method of calculation of the formants is described in Fig2.. Firstly, we get the cepstrums and the F_0 from the speech with autocorrelation method.

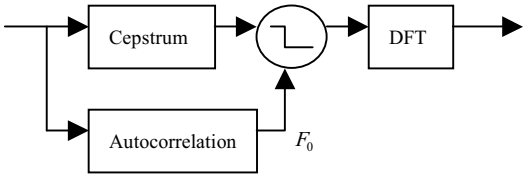

Fig2. The method in the calculation of formant spectrum

Because low part of the cepstrum of the speech contains vocal tract information, the formant can be got through the DFT precessing of the low part of the cepstrum.

Fig3 shows the result of the formant spectrogram got from the same sentence as for wide-band spectrum.

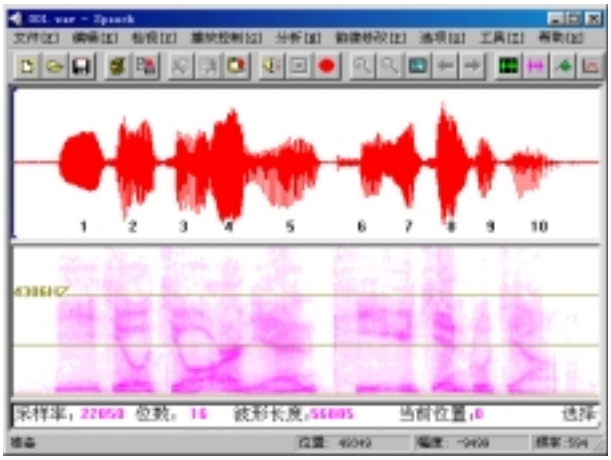

Fig3. Formant spectrogram

The method in getting the zero crossing rate of the speech is much simpler and the formula of it is described as following.

$$Z(t) = \sum_{m=-\infty}^{\infty} \left|\text{sgn}[s(m)] - \text{sgn}[s(m-1)]\right| \cdot h(t-m) \qquad (1)$$

Figure 4 shows the results of the zero crossing rate.

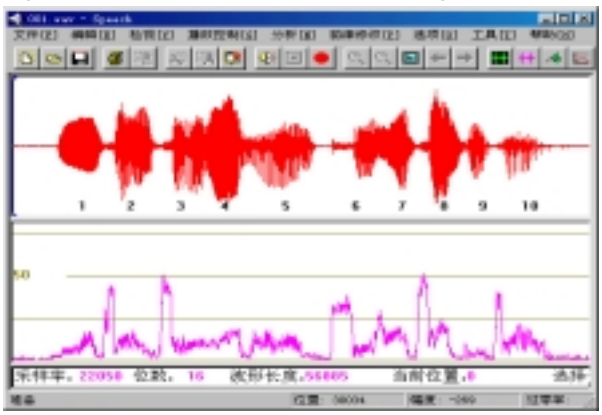

Fig4. Zero crossing rate contours

The acoustic parameters got by the analysis functions will help us analyze the frequency features of the speech, and they greatly facilitate us in applying the database.

4. PROSODIC LABELING

The prosodic labeling task requires mapping a sequence of feature vectors (mostly derived from the acoustic) to a sequence of labels. Although there are many features to be concerned, they are all derived from three basic acoustic features: duration, energy and F_0. Thus, in the tool, the prosodic labeling contains energy labeling, syllable segmenting and pitch labeling. In some special situation, such as for PSOLA method, the speech should be processed pitch synchronously. Thus, the pitch synchronous marks should also need to be labeled.

$F(\omega)$

Energy information of the speech database is influenced not only by the prosodic features of the speech but also by the recording environment. It's one of the important prosodic parameters.

The fomula of energy calculation is:

$$E(t) = \log 10[\sum_{m=-\infty}^{\infty} s^2(m) \cdot h(t-m)] \quad (2)$$

Figure 5 shows a result of energy contours of a speech.

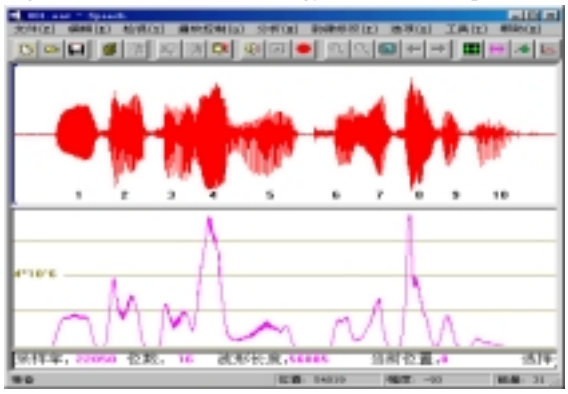

Fig5. Energy contours

Traditionally, the method of syllable segmenting is mainly based on the results of the energy $E(t)$ and the zero cross $Z(t)$ of the speech [3][7]. The boundary of the syllable is usually got through setting the threshold on the energy or zero cross parameters. With this method, the accuracy of the results is somewhat limited. Here, we proposed the formula 3&4.

$$R(t) = E(t) / Z(t) \quad (3)$$

$$X(t) = E(t) \cdot Z(t) \quad (4)$$

We replace E(t) and Z(t) with R(t), X(t) . With the analysis of X(t), the boundary of the most syllables are labeled. The accuracy of the results is improved to a certain extent. The boundaries between the voiced and unvoiced parts of the speech are also able to be detected by the analysis of R(t). The results are shown in the following figure.

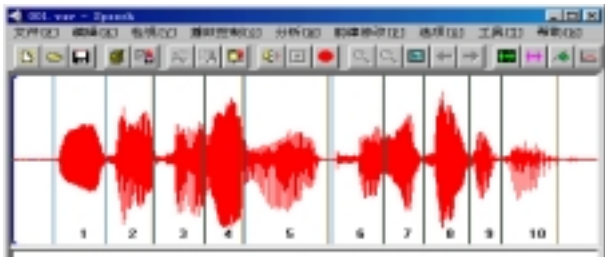

Fig6. Syllable boundaries labeling

Though the accuracy of the segmentation is improved, it is still difficult in divide the syllable in some special places. In such situation, the spectrum of the speech must be considered. After the syllable segmenting, the duration of the syllables and the silence between two syllables can be get easily.

In the tool, pitch labeling contains three parts: calculation and displaying of pitch contour of the speech, marking the glottal closure position and the period starting position of the speech, which can be thought to able to meet various needs of the users. The pitch contour is got with autocorrelation method [4] (results are shown in figure 8). But, the glottal closure positions and the period starting positions are much more difficult to be marked. Here we developed an automatic detecting method [2][5][6], which is shown in figure 7.

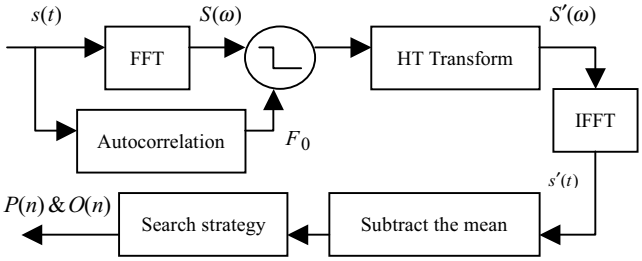

Fig 7. The method of marking the glottal closure positions and the starting positions

Here, the HT transform denotes a transform in frequency of the speech. The formula is as following,

$$HT(w) = \begin{cases} -j & 0 < w < \pi \\ 0 & w = 0, \pi \\ j & -\pi < w < 0 \end{cases} \quad (5)$$

$P(n)$ & $O(n)$ denote the glottal closure positions and the period starting positions of the speech. Figure 9&10 show the marking results separately.

In order to facilitate users, some of the labeling results,

such as syllable boundaries, the glottal closure positions and the starting positions, are allowed be adjusted manually by the user.

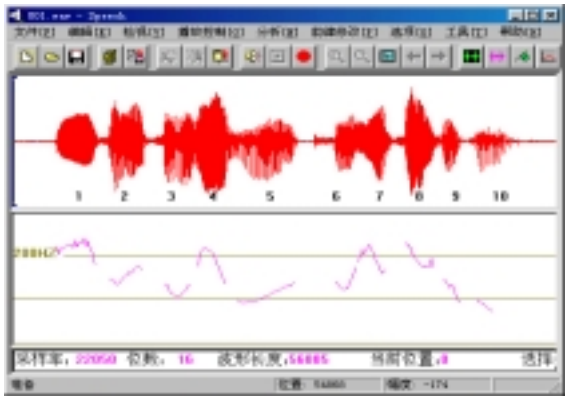

Fig8. Pitch contours

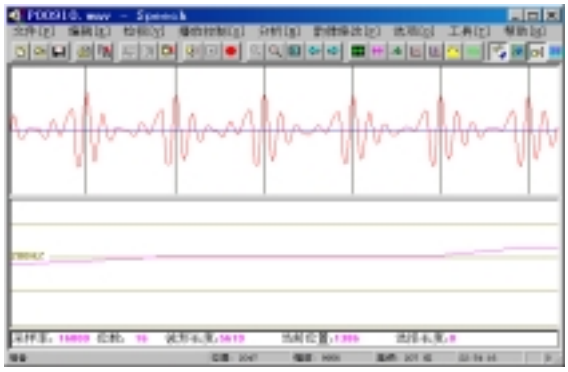

Fig9. The glottal closure positions

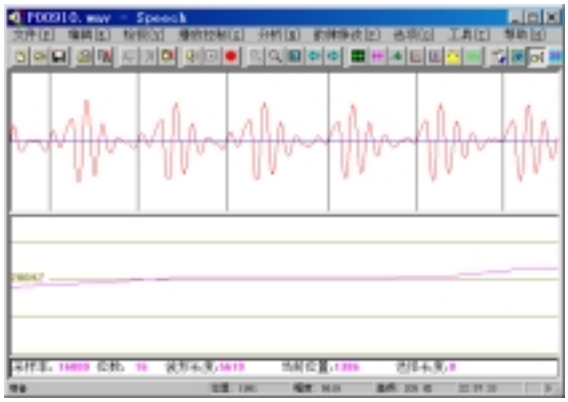

Fig10. The period starting positions

5. SPEECH PROCESSING

In order to apply the database in speech synthesis region, we particularly add some useful functions into the tool, which include: showing the pitch range of the syllable, modifying the volume or the pitch of the speech, filtering speech with Chebychev or Butterworth filter, etc. In figure 11, we can see that the pitch ranges are marked on. The F_0 contours can be modified, and the syllable can be resynthesized with the modified F_0 contours. It will be greatly help us to analyze prosodic characters of the speech and establish the prosodic model for the synthesis system.

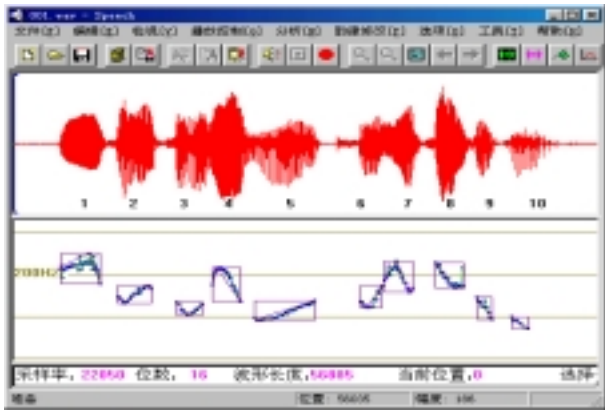

Fig11. Pitch ranges of ths syllables

The LP-PSOLA method is used in resynthsis processing, which is described in figure 12. The residual parameters e(t) got from the speech are modified using nonlinear phase modification method [1]. It can be thought that the method is superiorer than the traditional TD-PSOLA method.

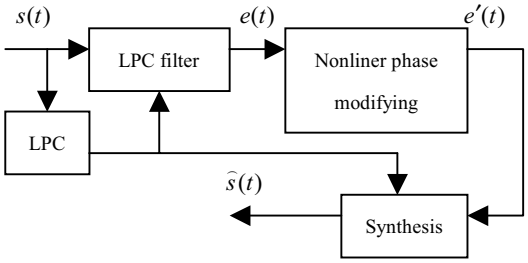

Fig11. Resynthesis method

6. APPLICATIONS AND RESULT

The tool can be easily operated. Now, it has been used widely in almost all aspects of our speech analysis and synthesis research. Especially, it has been used successfully in the labeling of speech synthesis database supported by National 863 High Technology Project.

The future version of the tool will be emphasis on the improvement of the method in parameter analysis and the addition of the new functions, which will be more suitable in improving the quality of the speech synthesis system. Moreover we'll take into account the functions, designed by us, could be shared by users directly.

REFERENCE

[1]Quatieri, T.F. and McAulay, R.J. (1992)," Shape

invariant time scale and pitch modification of speech", IEEE Transactions on Signal Processing, March 1992, VOL.40, pp 497-510.

[2]Tao Jianhua and Hua Yiman," Automatic Accurate Determining the Instants of Maximum Excitations in Voiced Speech", Chinese Journal of the applied acoustic, Oct, 1997, pp 21-25.

[3]L.R.Rabiner and R.W.Schafer," Digital processing of speech signals", Scientific publishing company, 1983.

[4]Y.Medan, E.Yair and D.Chazan," Super resolution pitch determination of speech signals", IEEE ASSP, 1991, VOL. 39, pp 40-47.

[5]Y.M.cheng, D.O'shaughnessy," Automatic and reliable estimation of glottal closure instant and period", IEEE ASSP, 1989, VOL. 37, pp 1805-1815.

[6]B.Yegnanarayana, R.L.H.M.Smits," A robust method for determining instants of major excitations in voiced speech", ICASSP, 1995, pp 776-779.

[7]Yang Xingjun and Chi Huisheng," Digital processing of speech signals", Electronic industry publishing company, 1995