# A Method for Decomposing and Modeling Jitter in Expressive Speech in Chinese

*Lei WANG*[1], *Aijun LI*[2], *Qiang FANG*[2]

[1] Dept. Computer Science,
Tianjin University, China
Willdx99j3@hotmail.com

[2] Institute of Linguistics,
Chinese Academy of Social Sciences
liaj@cass.org.cn

## Abstract

Jitter is considered as one of the most crucial factors to the aim of synthesizing natural motional speech. Unlike the traditional methods of measuring jitter in emotional speech, this paper propose that the jitter in the speech could be decomposed into two parts, that to say, deterministic jitter and random jitter. Deterministic jitter is associated with certain causes that may be the affect caused by emotion state, while random jitter is the result by random events that have nothing to do with emotion. What is more, two different methods of modeling jitter distribution are described: jitter decomposition is based on the fact that the mixed jitter can be divided into deterministic part and random part, while the algorithm based on GMM tries to simulate the shape of the histogram of jitter distribution. The result makes a qualitative analysis of the two methods. There are still much of works for us to do in the future in order to do more detail analysis and to make quantitative analysis of them.

## 1. Introduction

In the daily communication, the utterances not only carry the literal meanings, but also convey other information, such as emotion, attitude and so on [1]. The non-linguistic message is also a critical method to provide information about the speaker's mental state and intent beyond that disclosed by word content. For example, with the same literal meaning, listeners could perceive different understandings, even the opposite ones, due to the speakers' particular emotional purpose brought by the paralinguistic information [2]. With the development of speech processing technology, the study of this subject captivates more and more researchers' attentions [14, 18]. Current studies mainly cast focus on acoustic features and the most frequently used parameters is F0 contours, F0 range, jitter, stress frequency etc [2]. In this paper, we attempt to investigate emotional speech through the jitter modeling aspect.

Jitter is the deviation from the ideal timing of an event, and it is composed of both deterministic and Gaussian (random) content [3]. This concept is adopted by many fields, such as 1) in the computer networks, jitter is the variation in delay times experienced by the individual packets making up the data stream [4]. 2) In the electronic engineering field, Jitter is the deviation in or displacement of some aspect of the pulses in a high-frequency digital signal [5]. 3) In the phonetics, jitter involves small fluctuation of the glottal cycle lengths [6] and in this paper is defined as the deviation from the expected F0 contour. Jitter is very important in the area of speech synthesis, since if there is no jitter in the synthesized speech, it sounds like a machine sound in perception [8]. In this paper we assume that jitter distribution is determined by the speaker's emotional state, at least to some extent, and in addition, each emotion has its own jitter distribution. Therefore modeling the distribution of different emotions could be helpful to the emotional speech synthesis. However, within the former works, jitter is considered to be a Gaussian model, or normal model, which is suitable for modeling the pure random process (white noise) [8]. In this paper, we regard jitter distributing model as a distinctive feature to peculiar emotions, and try to modeling each of them.

There are a great many causes affecting the distribution of jitter. From emotional aspect, it may be the effect of tempestuously change in pitch, the alteration in voice quality, the stress pattern, etc. Those acoustic alterations can be realized by controlling of the configurations of thyroarytenoid muscle, changing of volume velocity of airflow, remodeling the surface of vocal tract, etc [9]. Nevertheless, nearly all the articulatory apparatus of a human being could be the source of jitter, and what we perceived is the result of mixed effects among those jitter's sources. Understanding the type and amount of jitter introduced by each emotion is basic and crucial process for further research of emotional speech synthesis and emotional recognition.

In the real world application, what is measured, in general, is the total effect of signal, which is mixed with the deterministic part and random part together. So does the jitter. To understand the underlying cause of the jitter, separating and identifying each jitter component is essential [7]. We are attempting to decomposing the total jitter(TJ) into two parts: the random jitter (RJ), which has nothing to do with emotions, only the random subtle variability in the speech, and the deterministic jitter(DJ), which is decided by emotion. Deterministic Jitter is created by identifiable interference signals. It is always bounded in amplitude and has specific non-random causes (here, we mainly focus on the emotional factors) [7]. In contrast to DJ, random jitter may be the result of accumulation of small random processes including asymmetry of right and left vocal fold, the influences of uneven distributed mucus on the surface of vocal fold, the turbulence caused by air through the glottis, and so on. Due to Law of Large Numbers, RJ can be represented by a Gaussian distribution which is characterized by its standard deviation (the mean value is obviously zero).

The motive of separating jitter is to find Jitter distribution pattern for different emotion. Since certain emotion associates with particular jitter distribution, computer can match the pattern to recognize emotion in speech. Further more, it can also be adopted by emotional speech synthesis to make the synthesized utterance more expressive.

This paper is composed of seven major parts. Section 2 introduces the emotional corpus we used, the way of classifying emotions, and a new method of quantify jitter. In section 3, the paper compares the statistics of jitter of different emotions. Section 4 proposes methods of how to modeling

jitter: separation jitter into DJ and RJ, GMM-based method of simulating the histogram of jitter distribution. A further analysis and discussion is given in section 5. Finally, a conclusion is made in section 6.

## 2. Corpus and jitter measuring

### 2.1. Classification of emotion

Emotion is usually considered to be a feeling about or reaction to certain important events or thoughts. It is one of our most difficult tasks to classify emotions. Psychologists have attempted to offer general classifications of these responses, and as with the color spectrum, systematically distinguishing between them largely depends on the level of precision desired. One of the most influential classification approaches is that of six primary emotions: anger, fear, sadness, disgust, surprise, and joy.

### 2.2. Corpus

To create the model, a corpus which contains 120 sentences and covers the most common Mandarin tone combination was designed. Each sentence contains about 7 syllables, and was recorded in 5 emotions, “neutral”, “joy”, “sadness”, “fear”, “anger”, by 2 professional actors and 2 professional actresses. The emotional sentences were recorded in the professional recording studio, and the emotional speech is elicited in this way: take sad emotion as an example, we first let the actor/actress read a very sad story, and then make he/she read the designed sentence in sad emotion in order to make the recorded speech as nature as possible.

### 2.3. Jitter measuring method

There are a number of jitter computing methods, but some of them are designed to estimate the whole effect of local jitter, so we must provide a more subtle way to quantify jitter. From the original definition of jitter, we adopt the following computing method for measuring jitter:

a, Using a Curve smoothing method to smooth the F0 contour, and take the smoothed F0 contour as the expected contour.

b, Employing an interpolating method to minimize error.

c, Because the pitch point extracted by the computer is sometimes inaccurate, a manually modifying method is needed.

d, Then we subtract the result of second step from the third step, and the residue is considered to be the deviation from the ideal position, that is to say: jitter.

e, Since there is a correlation between the subtracted result and the value F0, we measure jitter with a relative way: divide the expected value with subtracted result, then multiply with 100.

## 3. Statistical difference of emotions

We use the jitter measuring method above to calculate the jitter distribution of the five emotions, and the result histogram is shown as following:

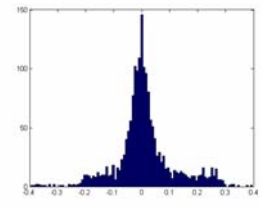

Figure 1: *fear*

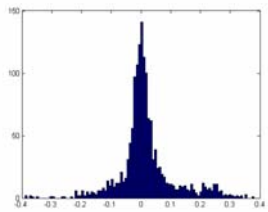

Figure 2: *sadness*

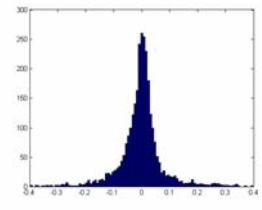

Figure 3: *anger*

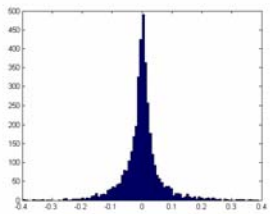

Figure 4: *joy*

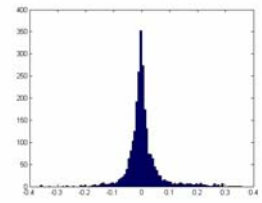

Figure 5: *neutral*

From these figures, we could get a general idea about the jitter distribution that fear and sadness have greater jitter variance than any other emotions, and neutral has the least jitter variance among these emotions.

## 4. Modeling of Jitter

As shown in the figures above, we can get the conclusion that the jitter distribution is a multimodal distribution. The multimodal distribution is resulted from several independent variables. That is to say, when there is a peak in the multimodal distribution, there must be a reason for it [7]. In this paper, jitter is decomposed to two characteristic components which are termed “deterministic” and “random”. The former is caused by systematic and emotional dependent sources, while the latter is associated with noise mechanisms [7].

Many methods have been proposed by the former researchers, [7] adopted the Tail fitting method, which requires little computation. Sompong Wisetphanichkij, Kobchai Dejhan used the derivatived Gaussian wavelet transform [10]. All those algorithm reached reasonable results. In this paper we employ two methods to separate and model the mixed jitter. One is based on multi-delta function and another is based on Gaussian Mixture Model.

### 4.1. The decomposition of jitter

From the histogram of jitter above, random jitter is assumed to be a Gaussian function with mean 0 and standard deviation $\sigma$ .

$$\mathrm{RJ}=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}} \qquad (1)$$

Deterministic jitter is assumed to be a multi-delta model [7], and the position of delta function can be considered as the influence of emotional state.

$$\text{DJ}=\sum_{i=1}^{N}\pi_i\delta(x-u_i),\ \sum_{i=1}^{N}\pi_i=1 \qquad (2)$$

From the Probability theory, the probability density of a sum of random processes is the convolution of their probability density functions, TJ=DJ*RJ. Thus the result of the convolution of the two PDF is

$$\sum_{i=1}^{N}\pi_i\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-u_i)^2}{2\sigma^2}}. \qquad (3)$$

The parameters in (3) can be evaluated by the maximum likelihood estimation.

### 4.2. Modeling jitter using the GMM

Gaussian Mixture Model (GMM) is a type of density model which comprise a number of weighted Gaussian component functions. These component functions are combined to provide a multimodal density and can model most of the real world probability models. The model assumes that the probability of observed parameters to take the following form:

$$\text{G(x)}=\sum_{i=1}^{N}\pi_i G(x\mid c_i),\ \sum_{i=1}^{N}\pi_i=1,\ \pi_i\geq 0 \qquad (4)$$

Where N is the numbers of Gaussian components: $G(x\mid c_i)$ is the ith Gaussian component and $\pi_i$ is the weight of ith component [11]. This model's parameter can be estimated by the Expectation Maximization (EM) algorithm. The expectation maximization (EM) algorithm is an iterative method for calculating maximum likelihood distribution parameter estimates from incomplete data (elements missing in feature vectors) [12].The algorithm can also be used to handle cases where an analytical approach for maximum likelihood estimation is infeasible, such as Gaussian mixtures with unknown and unrestricted covariance matrices and means.

### 4.3. The relationship between GMM & Jitter decomposition.

From the analysis above we can see that GMM is, to some extent, the same as the jitter decomposition in this study, because both of them are describing the same phenomena: the position and the shape of peaks in the histogram. Therefore these methods all intent to find out what is the emotion influence on jitter and how it is affected. With the enough data, the result of the two methods should be consistent as well.

## 5. Result and Discussion

### 5.1. The statistical property of jitter

We calculated the mean and variance value of those jitter distributions and the statistical property of jitter in different emotion is shown as Table 1:

Table 1: *statistical of jitter*

| | Neutral | Fear | Sad | Anger | joy |
|---|---|---|---|---|---|
| Mean | 0.00074 | 0.00054 | 0.0019 | -0.0046 | -0.0025 |
| Var. | 0.053 | 0.072 | 0.073 | 0.061 | 0.056 |

From the table 1, we could conclude that the mean value of jitter is very close to zero, and the sorted result shows that the sadness has the highest of variance value, while the neutral owns the least one. This result is consistent with the one obtained by the intuitionistic observation in Figure 1 to 5.

### 5.2. The result gained by GMM

The result of GMM by using 3 modals

Table 2: *GMM parameters of Neutral*

| Mean | Var. | Weight |
|---|---|---|
| 0.002 | 1.64E-04 | 0.4675 |
| -0.0051 | 0.001 | 0.4052 |
| 0.0148 | 0.0179 | 0.1273 |

Table 3: *GMM parameters of Fear*

| Mean | Var. | weight |
|---|---|---|
| -0.0209 | 0.0136 | 0.2567 |
| 0.0012 | 4.15E-04 | 0.7228 |
| 0.2459 | 0.0011 | 0.0205 |

Table 4: *GMM parameters of Neutral*

| Mean | Var. | weight |
|---|---|---|
| 2.72E-04 | 1.65E-04 | 0.4612 |
| -0.0018 | 0.0013 | 0.329 |
| 0.0123 | 0.0234 | 0.2097 |

Table 5: *GMM parameters of Anger*

| Mean | Var. | weight |
|---|---|---|
| -0.0213 | 0.002 | 0.3273 |
| -0.0168 | 0.0215 | 0.126 |
| 0.0082 | 3.27E-04 | 0.5467 |

Table 6: *GMM parameters of Joy*

| Mean | Var. | weight |
|---|---|---|
| -0.0131 | 0.0021 | 0.3314 |
| -0.0038 | 0.0184 | 0.1091 |
| 0.0041 | 1.97E-04 | 0.5595 |

### 5.3. The comparing of five emotions

There are three parameters within the result of GMM model: Mean, variance and weight. Mean and variance value denote the position and dispersion of corresponding Gaussian distribution, and the value of weight reflects the height of this peak. From the result of GMM modeling, we could also reach this result that among all of the emotions we measured, sadness is the most unstable one, while gladness and neutral is relative stable.

### 5.4. The RJ and DJ in emotional speech

As shown in the Table 2 to 6, we could conclude that the middle peak of GMM parameters are very similar among different emotions; it may be caused by random jitter. The main differences among those emotions' parameters are the left and right peaks; it may be caused by the different emotions, that is to say, deterministic jitter in our definition.

### 5.5. The factors may influence the distribution of jitter

Ingo R. Titze in his paper [16] concludes that there are numerous factors that could bring on the various jitter styles in the speech, such as the genetics, development and aging, language and culture, disease, food and drug, vocal training, and emotions.

In addition, Chinese is a tonal language, there are 4 tones in Chinese Mandarin, and each tone has its own F0 and range. We attempt to eliminate the effect of tones by carefully designing the sentences; nevertheless, there is still a few of remaining. Since the jitter's value is associated with the value of pitch, the jitter distribution could be influenced by the different tone considerably.

What is more, some special voice quality account for the uneven distribution of jitter. For example, the creaky voice often occurs in the context of tone 3 in Mandarin, in this case, jitter is more even than other voice qualities; breathy voice often happens when people want show their intimacy to others, and in this situation, the jitter distribution is very different with others. The measuring of distribution of different voice qualities is also another main work for us to do at present and we will get some interesting result soon. Also, the detailed shape of the histogram of jitter distribution maybe different depending on the speakers, however, the pattern of the histogram remains the same in each emotion.

Last but not the least, emotions are crucial factors affects the jitter distribution. Emotion is the attitude generated from the brain, and encoded with the physiological organ, then transmitting with the sound waves. Different emotions may have their own patterns on phonation apparatus, that is to say: the left and right vocal fold, on the false vocal folds, subglottal and supraglottal tract, airflow to induce dissimilar jitter distributions. In addition, even heart beat and mucus on the surface of vocal folds caused by emotions could be factors owing to the jitter distribution.

### 5.6. The factors may affect the result

First of all, in noisy environments, it is difficult to estimate accurate F0 because of the interference of noise. Although we try to avoid the noise in our recording studio, there are still some brought by the unavoidable factors, such as alternating current, people's breathy and other trivia. This could be one of the chief reasons for the inaccuracy.

Moreover, the corpus we used in this paper is in Chinese. Chinese is a tonal language, and tones may have their peculiar jitter distribution. This influence can not be underestimated.

In addition, the data size is not enough, thus there may be errors in the quantitative analysis. However, the qualitative analysis might reflect the situation in the real world.

## 6. Conclusion

This paper analyses the emotion influences on jitter distribution, and proposes that there are two components mixed in the jitter we perceived in the real world. Then two different method of jitter modeling are described: jitter decomposition is based on the fact of total jitter can be divided into deterministic part which is caused by the specific emotions and random part caused by the random events during the speech production, while the algorithm based on GMM tries to simulate the shape of the jitter distribution.

From the result gained from jitter data, we compared the different jitter distribution of different emotions, and it is consistent with the people's experience. However, it is just the beginning of this work; there are numerous works left to be done in the further studies.

## 7. Acknowledgement

This research was funded by the National Science Foundation, Project No.60275015.

## 8. References

[1] N. Campbell, "Towards a grammar of spoken language: Incorporating paralinguistic information," in ICSLP2002, Denver, CO., 2002.

[2] 陶建华, 听话要听音—情感语音处理技术 www.ctiforum.com/forum/2005/02/forum05_0205.htm

[3] T11.2 / Project 1230/ Rev 10 Fibre Channel - Methodologies for Jitter Specification page 7.

[4] Mansour J. Karam, Fouad A. Tobagi: Analysis of the Delay and Jitter of Voice Traffic Over the Internet. INFOCOM 2001: 824-833

[5] Masashi Shimanouchi, "An Approach to Consistent Jitter Modeling for Various Jitter Aspects and Measurement Methods", IEEE International Test Conference, pp 848 - 857, 2001.

[6] Lieberman, P. "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges," J. Acoust. Soc. Am. 35, 344–353.

[7] Mike P. Li, Jan B. Wilstrup, Ross Jessen, Dennis Petrich: A new method for jitter decomposition through its distribution tail fitting. ITC 1999: 788-794

[8] Xiyu WU, 汉语普通话嗓音声学研究, master thesis of Chinese Academy of Social Science, 2004.

[9] K. H. Kim, S. W. Bang, S. R. Kim, Emotion recognition system using short-term monitoring of physiological signals, Medical & Biological Engineering & Computing 2004, Vol. 42, pp. 419 – 427, 2004

[10] Sompong Wisetphanichkij, Kobchai Dejhan, Jitter Decomposition by Derivatived Gaussian Wavelet Transform, The 2004 International Symposium on Communications and Information Technologies.

[11] http://homepages.inf.ed.ac.uk/rbf/CVonline/LOCAL_COPIES/RAJA/CV.html

[12] http://www.cs.brown.edu/research/ai/dynamics/tutorial/Documents/ExpectationMaximization.html

[13] Zetterholm, E. 1998. 'Prosody and voice quality in the expression of emotions'. Proceedings of the Seventh Australian International Conference on Speech Science and Technology, 109-113. Sydney, Australia.

[14] J. Schoentgen, "Stochastic models of jitter" J. of Acoust. Soc. of America, April 2001, vol. 109, no. 4, pp. 1631−1650.

[15] J. E. Cahn. The generation of affect in synthesized speech. Journal of the American Voice I/O Society, 8:1-19, July 1990.

[16] I. R. Titze, "Workshop on Acoustic Voice Analysis, Summary Statement," National Center for Voice and Speech, Denver,. Colorado, 1994.

[17] Cécile Pereira, Perception and Expression of Emotion in Speech, doctor thesis of Macquarie University, 2000.

[18] Schoentgen, J et De Guchteneere, R (1995)Time-seriesanalysis of jitter. Journal of Phonetics, 23-1, pp 189-201.