Signal processing: New Stochastic Feature of Unvoiced Pronunciation for Whisper Speech Modeling and Synthesis

X. D. Zhuang[1,3], H. Zhu[2], N. E. Mastorakis[3]
[1]Electronic Information College, Qingdao University, 266071 China (xdzhuang@qdu.edu.cn)
[2]Qingdao University of Science and Technology, 266061 China
[3]Technical University of Sofia, Industrial Engineering Department, Kliment Ohridski 8, Sofia, 1000 Bulgaria

Abstract—Whisper is an indispensable way in speech communication, especially for private conversation or human-machine interaction in public places such as library and hospital. Whisper is unvoiced pronunciation, and voiceless sound is usually considered as noise-like signals. However, unvoiced sound has unique acoustic features and can carry enough information for effective communication. Although it is a significant form of communication, currently there is much less research work on whisper signal than common speech and voiced pronunciation. Our work extends the research of unvoiced pronunciation signal by introducing a novel signal feature, which is further applied in unvoiced signal modeling and whisper sound synthesis. The statistics of amplitude for each frequency component is studied individually, based on which a new feature of "consistent standard deviation coefficient" is revealed for the amplitude spectrum of unvoiced pronunciation. A synthesis method for unvoiced pronunciation is proposed based on the new feature, which is implemented by STFT with artificially generated short-time spectrum with random amplitude and phase. The synthesis results have identical quality of auditory perception as the original pronunciation, and have similar autocorrelation as that of the original signal, which proves the effectiveness of the proposed stochastic model of short-time spectrum for unvoiced pronunciation.

Keywords—Signal Processing, whisper, unvoiced pronunciation, short-time spectrum, speech synthesis, standard deviation coefficient

I. INTRODUCTION

WHISPER is a socially significant form of daily speech communication. It is characterized by the lack of vocal cord vibration, which implies the absence of fundamental pitch. Current development of speech technology introduces new leading edge areas where whisper becomes of interest [1]. It becomes more popular due to the widespread use of smart phones and pads. Current research hotspots include text-to-whispered-speech (as the case of the Whisper Mode of Amazon Alexa), whisper speech recognition, speaker identification in whisper speech [2-6]. These technologies will be possibly promoted significantly if whisper signal feature be well studied in the perceptual, acoustical, and signal analysis domains [7-12]. However, despite the abundant research literature supporting normal speech, relatively little effort has been spent on whisper speech, especially the study on unique features of whisper speech in perceptual and signal analysis domain [13-15]. Although there has been sufficient work on normal pronunciation synthesis, there is also a need for efficient method to synthesize whisper speech for human-machine communication [16-18]. In his paper, we aim at study the unique or specific signal features of unvoiced sound in whisper speech, and attempt to give an efficient model of unvoiced pronunciation. We also attempt to provide an efficient framework for whisper pronunciation synthesis based on the signal features. To our knowledge, there is very little study attempting to achieve this goal before [19].

Whisper speech is of unvoiced pronunciation. Speech signal can be mathematically modeled by stochastic process, especially the unvoiced pronunciation in whisper. There have been researches on stochastic properties of continuous speech signals (i.e. signals of daily conversations). Such researches are based on the large amount of speech data in corpora like TIMIT [20], AURORA [21] or other database of daily speech signal from the internet [22]. These studies have investigated the probability distribution for time-domain speech signal, and also for the data in transformed domain as well, such as DCT, KLT, DFT, etc. In these studies, several probability distributions have been tested to propose a stochastic model for comprehensible speech with meaningful language contents (i.e. sentences or paragraphs). Such distributions include the Gaussian distribution (GD), Laplacian distribution (LD), Gamma distribution (ΓD), Generalized Gaussian distribution (GGD), and Generalized Gamma distribution (GΓD). The probability distribution function (pdf) of time-domain speech signal was first studied in 1950s and 1960s, in which the Gamma and Laplacian distributions were tested [23,24]. The two-side Gamma distribution was also found to be a good approximation of the underlying pdf for time-domain speech samples [25-27]. For different speech lengths and different speech classes, the most proper pdf type was studied by Chi-square goodness-of-fit test on time-domain speech signal, with Laplacian and Gaussian distribution as two options [28]. Besides the

time-domain, the pdf for transformed domain was also studied for Karhunen-Loeve (K-L) transform and DCT, where the Laplacian distribution showed prevailing performance by Chi-square test and moment test [21]. In frequency domain, the pdf of the spectrum's real and imaginary parts was studied and modeled by Generalized Gaussian distribution based on Kullback-Leibler divergence as the fitting criterion [29]. And the amplitude spectrum of speech signal was modeled as Rayleigh distribution [30]. These stochastic models facilitates the application of speech enhancement [31-34] and voice activity detection [35], whose object is the daily-life speech interfered by some kind of noise. Suchmodels also facilitate speech coding [25] and speech recognition [36,37].

The stochastic property of speech signal can be analyzed on multiple levels. For one level, the speech signal in daily communication varies with the continuously changing of the language content (i.e. different words in sentences). This can be regarded as randomness at the language content level in speech. On the other hand, for the same language content (i.e. the same word or sentence), the speech signal will also vary randomly due to the speaker characteristics such as gender, age, emotion, etc. Moreover, even for a sustained pronunciation (such as a single phoneme) by a specific speaker, randomness still exists in the signal, which is caused by the physical mechanism of pronunciation. Such randomness has been observed in previous research. For voiced pronunciation there are phenomena of jitter and shimmer observed [38-40], and for unvoiced pronunciation the sound source is random by itself which is caused by turbulence of air flow in the vocal tract [41]. Such randomness can be regarded as another different level of speech randomness.

Although several probability models have been proposed for speech signal, the different levels of speech randomness discussed above have not been differentiated and studied separately. Since these studies are based on the large amount of speech data as daily-life sentences in corpora like TIMIT, AURORA, etc., the current stochastic models eventually represent the overall stochastic property, which is the combination of different randomness levels in speech mentioned above. However, the overall statistic property of speech can not represent the specific properties of different pronunciation types, which is important for deeper understanding of pronunciation mechanism, and also improvement of algorithms in practical applications. As a fundamental aspect for understanding the nature of speech signal, detailed stochastic properties of different specific types of pronunciation need to be studied. So far as we know, there is very little study on the stochastic distribution of short-time spectrum of specific unvoiced pronunciations yet, which may reveal intrinsic property of such speech pronunciation.

The unvoiced pronunciation is closely related to the aerodynamic process in vocal tract [41-46]. The physical process during unvoiced pronunciation is complicated, while the stochastic study of the signal produced may reveal some underlying properties of this process. The randomness in time-domain signal corresponds to the random fluctuation of its short-time spectrum. The focus of this paper is the random fluctuation of amplitude and phase for specific frequency component in the short-time spectrum of a sustained unvoiced pronunciation (or unvoiced phoneme), which forms the whisper speech. Moreover, the relationship between two different frequency components in amplitude probability distribution is also investigated, which is important but captured little research attention before.

The main contributions of this paper are: (1) a new feature of short time spectrum for unvoiced signal is revealed, which has not been reported before, and (2) a new efficient framework is proposed for whisper pronunciation synthesis, which is based on the signal model derived from the new spectrum feature.

II. New Feature for Amplitude Spectrum of Unvoiced Pronunciation

In the discrete spectrum obtained by STFT, we consider the spectrum value of each discrete frequency component as random variable due to the randomness of the signal. For each frequency component, we study the statistics of short-time amplitude spectrum by estimating the expectation and standard deviation as two basic statistics. A novel stochastic property about the relationship between these two basic statistics is revealed for unvoiced pronunciation. Because large amount of unvoiced data is needed for this statistic study, the signals captured and used in this paper are sustained unvoiced pronunciations, not the words or sentences in daily communication. The study concentrates on the signal randomness at the level of single unvoiced phoneme.

Although well-organized corpuses like TIMIT have well labeled the detailed words and syllables on the time axis, the single pronunciations in such data are too short for statistic study. Since currently there is little corpus of sustained phoneme pronunciation, signals have been captured by using microphones connected to the sound card on computers. To guarantee the generality of analysis, signals have been captured for a group of unvoiced pronunciation by different speakers, and on different recording platforms (different microphones and sound cards on different computers). In the collection of signal, the speakers were informed with the requirements of stable pronunciation for sufficient time length, based on which reliable statistic study can be achieved. For each speaker the signals were captured repeatedly for several times, so that the most stable signal suitable for study can be selected. The signals were recorded at sample frequency of 16 kHz, with 16 bit per sample. Because the unvoiced pronunciation is produced by the aero-acoustic process in the vocal tract without vocal cord vibration, it is much less affected by individual difference such as age and gender.

For a signal of a sustained unvoiced pronunciation, STFT is performed on that signal. Large amount of short-time spectrum data can then be obtained, based on which the expectation and variance can be estimated for each frequency component. In the experiments, the frame length is 512, which corresponds to a time interval of 32ms with a 16 kHz sampling frequency. A Hamming window is used on each frame for STFT. For all frequency components, the unbiased estimation of amplitude expectation and variance can be represented by two functions

$\mu(\omega_k)$ and $\sigma^2(\omega_k)$ respectively:

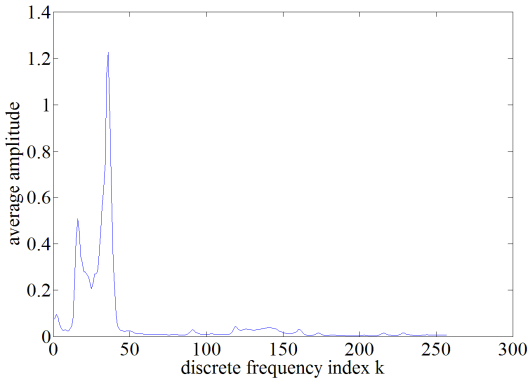
$$\mu(\omega_k) = \frac{1}{N}\sum_{i=1}^{N} a^i(\omega_k) \tag{1}$$

$$\sigma^2(\omega_k) = \frac{1}{N-1}\sum_{i=1}^{N}\left(a^i(\omega_k) - \mu(\omega_k)\right)^2 \tag{2}$$

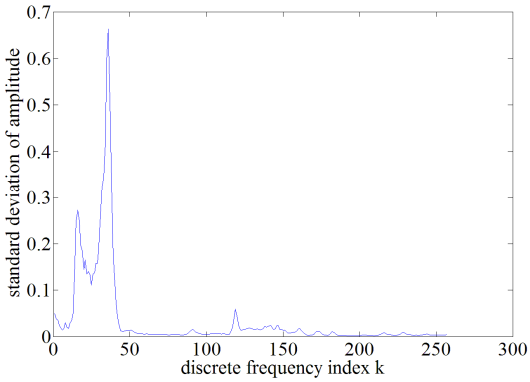
where N is the frame number, ω_k is the k-th frequency component in STFT, and $a^i(\omega_k)$ is the amplitude spectrum value of ω_k for the i-th frame. Moreover, the standard deviation $\sigma(\omega_k)$ is also estimated as the square root of $\sigma^2(\omega_k)$:

$$\sigma(\omega_k) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(a^i(\omega_k) - \mu(\omega_k)\right)^2} \tag{3}$$

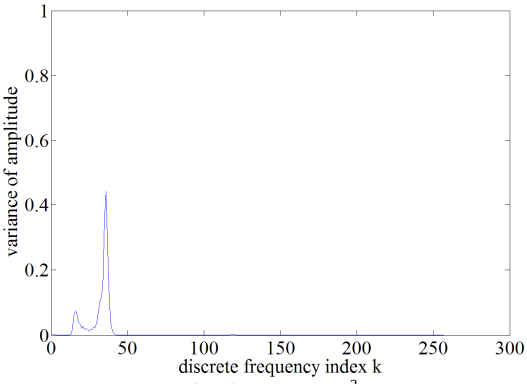
Some of the estimation results are shown from Fig. 1 to Fig. 4 (for the pronunciation of [h], [s], unvoiced [a], unvoiced [e] by a male speaker), where the curves of $\mu(\omega_k)$, $\sigma^2(\omega_k)$ and $\sigma(\omega_k)$ are plotted for comparison. By comparing the curves of $\mu(\omega_k)$ and $\sigma(\omega_k)$ in each figure, their evident similarity can be observed, which inspires further study of the relationship between $\mu(\omega_k)$ and $\sigma(\omega_k)$.



(a) amplitude expectation $\mu(\omega_k)$

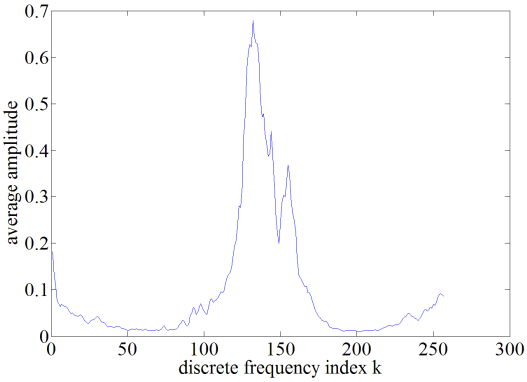


(b) amplitude standard deviation $\sigma(\omega_k)$

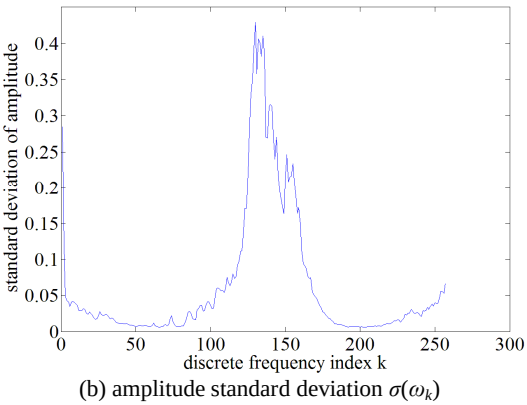


(c) amplitude variance $\sigma^2(\omega_k)$

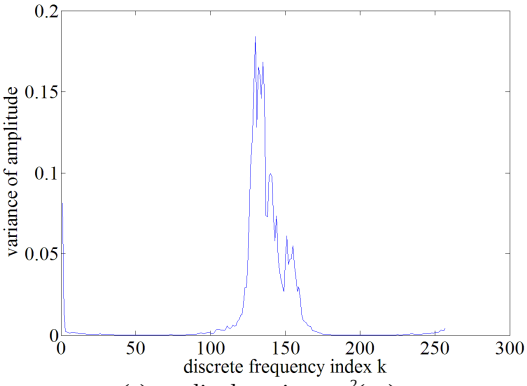
Fig. 1. The estimated expectation, variance and standard deviation of the short-time amplitude spectrum for [h]



(a) amplitude expectation $\mu(\omega_k)$

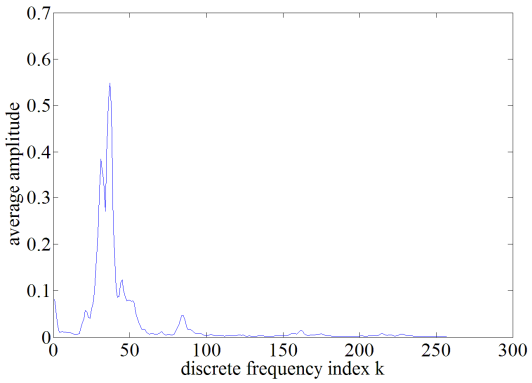


(b) amplitude standard deviation $\sigma(\omega_k)$

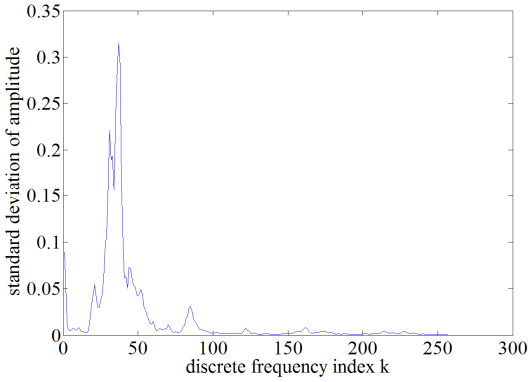


(c) amplitude variance $\sigma^2(\omega_k)$

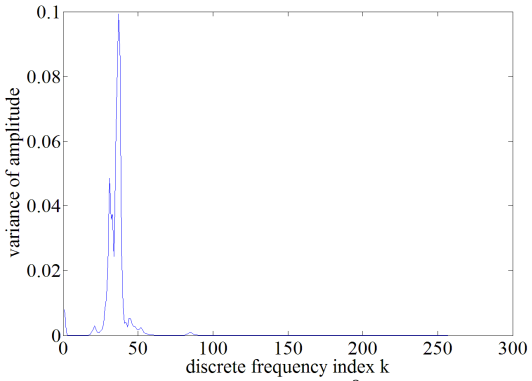
Fig. 2. The estimated expectation, variance and standard deviation of the short-time amplitude spectrum for [s]



(a) amplitude expectation $\mu(\omega_k)$

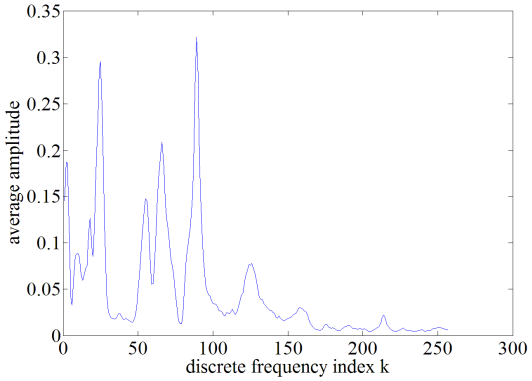


(b) amplitude standard deviation $\sigma(\omega_k)$

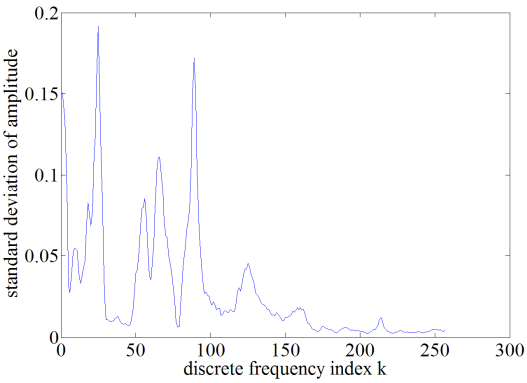


(c) amplitude variance $\sigma^2(\omega_k)$

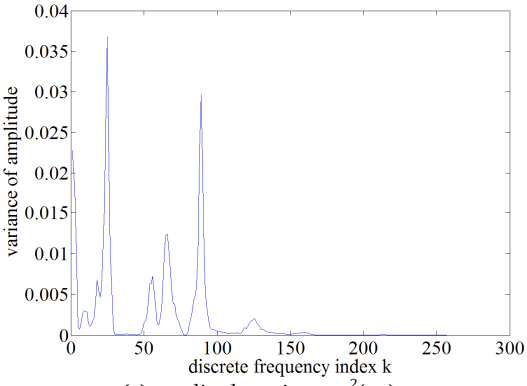
Fig. 3. The estimated expectation, variance and standard deviation of the short-time amplitude spectrum for unvoiced [a]



(a) amplitude expectation $\mu(\omega_k)$



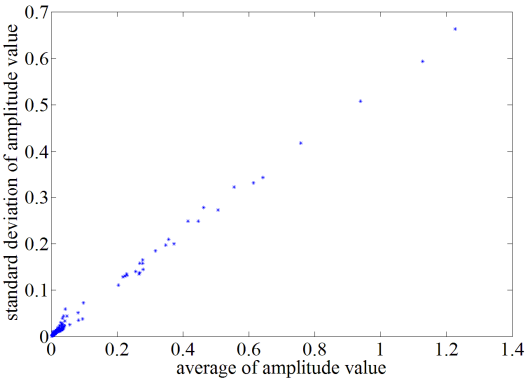
(b) amplitude standard deviation $\sigma(\omega_k)$



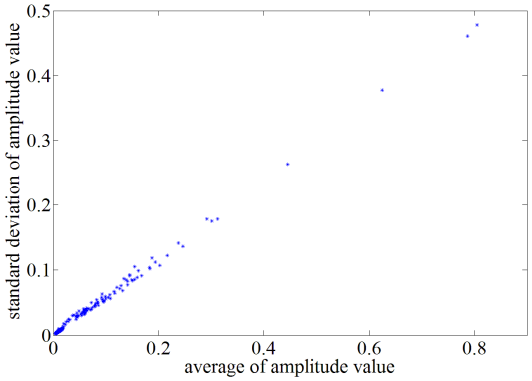
(c) amplitude variance $\sigma^2(\omega_k)$

Fig. 4. The estimated expectation, variance and standard deviation of the short-time amplitude spectrum for unvoiced [e]

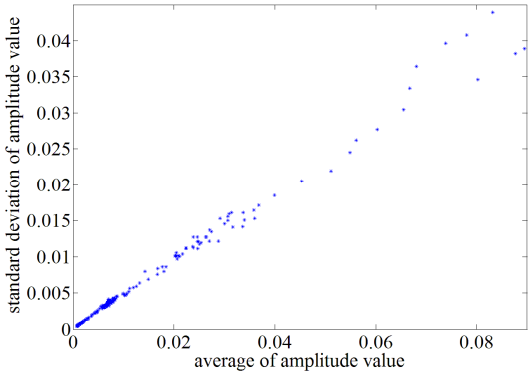
In order to investigate the relationship between $\mu(\omega_k)$ and $\sigma(\omega_k)$, for an unvoiced pronunciation, the two-dimensional points of $(\mu(\omega_k), \sigma(\omega_k))$ are plotted in Matlab for all ω_k. For a frequency component ω_k, $(\mu(\omega_k), \sigma(\omega_k))$ is a point with the amplitude expectation as the x-coordinate and the amplitude standard deviation as the y-coordinate. The results of such plotting indicate a linear proportional relationship between $\mu(\omega_k)$ and $\sigma(\omega_k)$. Some results are shown in Fig. 5, which demonstrate the linear relationship between $\mu(\omega_k)$ and $\sigma(\omega_k)$ in a more direct way.



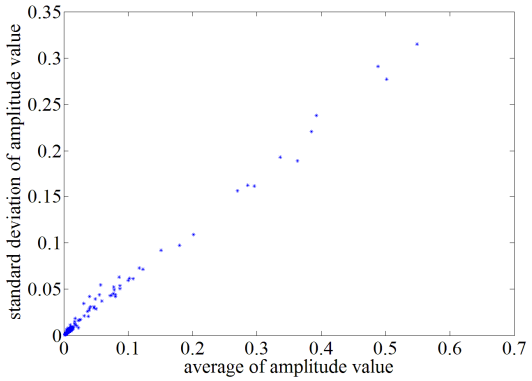
(a) The result for [h]



(b) The result for unvoiced [i]



(c) The result for [f]



(d) The result for unvoiced [a]

Fig. 5 The distribution of the points ($\mu(\omega_k)$, $\sigma(\omega_k)$) by point plotting in Matlab

Besides the above experimental results, the relationship between $\mu(\omega_k)$ and $\sigma(\omega_k)$ is quantitatively verified by calculating the correlation coefficient between the two curves of $\mu(\omega_k)$ and $\sigma(\omega_k)$. Since the spectrum obtained by STFT is discrete in frequency domain, the correlation coefficient is calculated in a discrete form:

$$\rho_{\sigma\mu} = \frac{\sum_{k=1}^{N} \sigma(\omega_k) \cdot \mu(\omega_k)}{\sqrt{\sum_{k=1}^{N} \sigma^2(\omega_k)} \cdot \sqrt{\sum_{k=1}^{N} \mu^2(\omega_k)}} \tag{4}$$

where N is the number of discrete frequencies in the discrete spectrum.

Experimental results are shown in Table 1. The first part of the results are based on the pronunciation signals recorded for one male speaker. The correlation coefficients between $\mu(\omega_k)$ and $\sigma(\omega_k)$ are calculated for different unvoiced phonemes, together with those between $\mu(\omega_k)$ and $\sigma^2(\omega_k)$ for comparison. The correlation coefficients between $\mu(\omega_k)$ and $\sigma(\omega_k)$ are close to 1.0. Consider the error caused by the instability of natural pronunciation, and also the noise introduced in the signal capture process, the strong correlation between $\mu(\omega_k)$ and $\sigma^2(\omega_k)$ observed in the experiments did not happen merely by chance. This results prove that $\mu(\omega_k)$ and $\sigma(\omega_k)$ are related by a linear proportional relationship.

Experiments have also been done on the unvoiced pronunciation recorded for other speakers, and by other recording devices, which also yield results indicating the linear proportional relationship between $\mu(\omega_k)$ and $\sigma(\omega_k)$. Some of these results are also shown in Table 1.

Since the standard deviation coefficient represents the σ to μ ratio, the new property revealed in the experiments is named as "consistent standard deviation coefficient". It means that the proportional coefficient between the standard deviation and the expectation is consistent for all the frequency components in the short-time spectrum of unvoiced pronunciation. From the viewpoint of probability, for any frequency component of an unvoiced pronunciation, the larger the amplitude expectation, the larger the random fluctuation of amplitude in the short-time spectrum..

Table 1 The correlation coefficient of $\mu(\omega_k)$, $\sigma(\omega_k)$ and $\sigma^2(\omega_k)$ for unvoiced pronunciation

Pronunciation	**ρ between $\mu(\omega_k)$ and $\sigma(\omega_k)$**	**ρ between $\mu(\omega_k)$ and $\sigma^2(\omega_k)$**	**Number of frames**
[s] (male)	0.9910	0.9375	35748
[θ] (male)	0.9852	0.8877	28126
[f] (male)	0.9948	0.8946	40179
[h] (male)	0.9982	0.9225	21909
unvoiced [a] (male)	0.9960	0.9374	17497
unvoiced [ə] (male)	0.9817	0.8933	45336
unvoiced [e] (male)	0.9913	0.9022	41872
unvoiced [i] (male)	0.9896	0.8525	44147
unvoiced [a] (female)	0.9960	0.9374	17497
unvoiced [ə] (female)	0.9817	0.8933	45336
unvoiced [a] (male; recorded on another audio capture platform)	0.9980	0.9130	39727
unvoiced [ə] (male; recorded on another audio capture platform)	0.9890	0.8978	21464

III. The Statistical Analysis of Amplitude Spectrum for Unvoiced Phoneme

Besides the basic statistics mentioned in Section 2, the histogram of amplitude value for each frequency component is also computed respectively using a voting method, which corresponds to the amplitude probability distribution. Then a new model is proposed as a uniform amplitude distributions for different frequency components. The new amplitude pdf model accords well with the property of "consistent standard deviation coefficient"..

A. The estimation of ω_k's amplitude distribution

For amplitude histogram computation, a voting method is

presented here. For each frequency component, the following steps are carried out:

Step 1: Determine a reasonable range of amplitude value for this frequency component. This range should contain all the amplitude spectrum values obtained from experimental data.

Step 2: Uniformly divide the above range into reasonable amount of intervals (or bins) with equal length.

Step 3: For each amplitude value of this frequency component, find the bin into which it falls. Increase the count of that bin by one (the voting).

After the above voting process, the amplitude histogram can be computed by dividing the voting results by the total number of data. In **Step 2**, the length of the interval or bin for the amplitude range can be determined by experiment. In order to conveniently compare the amplitude distribution for any two frequency components, a common value range $[0, A_{\max}]$ is used for all the frequency components, where $A_{\max}$ is the maximum of all the amplitude values for all the frequency components.

Besides the amplitude distribution of each frequency component alone, the connection between the amplitude probability distribution of any two frequency component is also important. To investigate that, the estimated distribution curves of all the frequency components are plotted together and shown as family of curves. In Section 2, the experimental results indicate that the expectation and variance of the amplitude spectrum value for different frequency components are usually different. The expectation and variance can determine the location and sharpness of the distribution curve. The shape of estimated amplitude distribution curve for each frequency component is affected by the corresponding expectation, which may make it inconvenient for study the connection between two amplitude probability distribution curves estimated.

Therefore, for each frequency component, a preprocessing step is added to normalize the expectation of amplitude values, so that the connection between the amplitude probability distribution of different frequencies may be revealed more clearly. Considering the linear proportional relationship between $\mu(\omega_k)$ and $\sigma(\omega_k)$, the preprocessing is proposed as dividing each amplitude spectrum data by the average amplitude value of its corresponding frequency component. This preprocessing is called "expectation-normalization" hereafter, because after such processing the data will have an average of 1. Some of the amplitude distributions after the preprocessing of expectation-normalization are shown in Fig. 6 to Fig. 11.

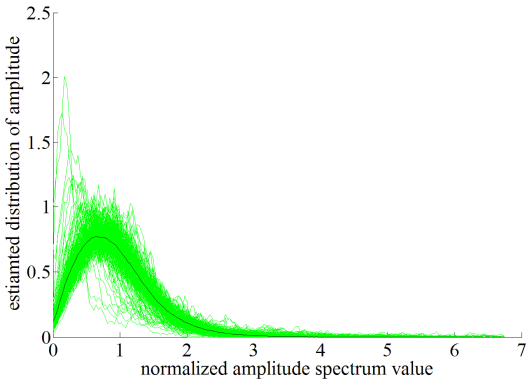


Fig. 6. The estimated amplitude probability distribution of each frequency ω_k for [h]

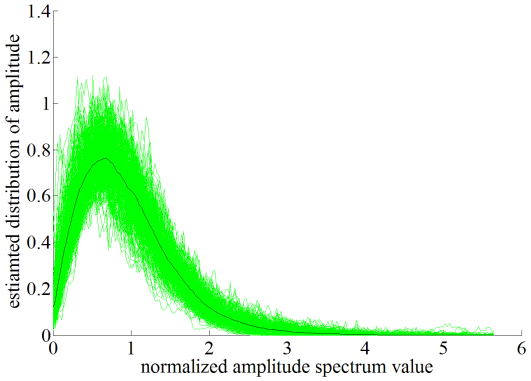


Fig. 7. The estimated amplitude probability distribution of each frequency ω_k for [s]

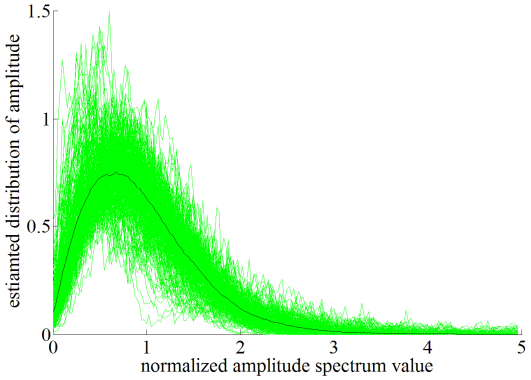


Fig. 8. The estimated amplitude probability distribution of each frequency ω_k for unvoiced [a]

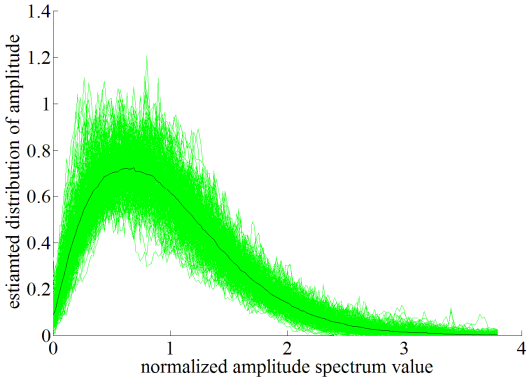


Fig. 9. The estimated amplitude probability distribution of each frequency ω_k for unvoiced [e]

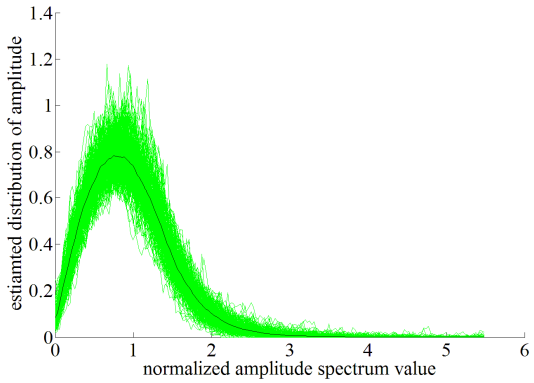


Fig. 10. The estimated amplitude probability distribution of each frequency ω_k for [θ]

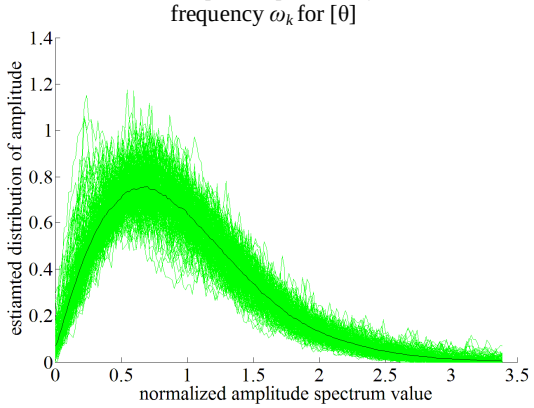


Fig. 11. The estimated amplitude probability distribution of each frequency ω_k for unvoiced [i]

The results show that, after the preprocessing of expectation-normalization, the estimated distribution curves obviously converge to one central curve (shown in black color in Fig. 6 to Fig. 11). Because the distribution curves converge so closely, the mixed plotting results in a belt around a central curve. From each figure, the strong connection between the amplitude distributions of different frequency components is indicated.

B. A model of the amplitude distributions for an unvoiced phoneme

In Fig. 6 to Fig. 11, the estimated distribution curves after expectation-normalization converge closely to one central curve. Considering the inevitable error caused by pronunciation instability and noise, it is reasonable to propose a common pdf prototype of amplitude for all frequency components in a single unvoiced phoneme. In another word, the amplitude distributions of different frequency components are of the same pdf type, but with different expectation values. In this model, there is a prototype distribution function $p_0(a_0)$, from which the amplitude distribution of any frequency component can be derived by varying the expectation (i.e. altering the expectation with a scaling factor). The prototype $p_0(a_0)$ corresponds to the central curve to which the estimated curves converge in Fig. 6 to Fig. 11. As a random variable, the amplitude of a frequency component a is modeled as some scaling of a prototype variable a_0, whose expectation is 1 (the unit expectation):

$$a = k \cdot a_0 \tag{5}$$

where k is the scaling parameter. Equation (5) is a mathematical description of the model proposed. Different ω_k may have different value of k, but a_0 is unique for each frequency component in a single unvoiced phoneme.

Moreover, we prove that this amplitude model accords well with the "consistent standard deviation coefficient" property. First, we derive the probability distribution of a in Equation (5), given the probability distribution of a_0 as the prototype distribution $p_0(a_0)$. The expectation of a is:

$$\mu_a = E[a] = E[k \cdot a_0] = k \cdot E[a_0] = k \cdot \mu_0 \tag{6}$$

where μ_0 is the expectation of a_0.

Based on the probability theory, the probability distribution of a can then be deduced as:

$$p(a) = \frac{1}{k} \cdot p_0\left(\frac{a}{k}\right) \tag{7}$$

Second, we derive the standard deviation coefficient of a:

$$\frac{\sigma_a}{\mu_a} = \frac{\sqrt{Var(a)}}{\mu_a} = \frac{\sqrt{\int_{-\infty}^{+\infty} (a-\mu_a)^2 \, p(a) da}}{\mu_a} \tag{8}$$

Consider Equation (6) and (7), Equation (8) can be rewritten as:

$$\frac{\sigma_a}{\mu_a} = \frac{\sqrt{\int_{-\infty}^{+\infty} (a - k\cdot\mu_0)^2 \cdot \frac{1}{k} p_0\left(\frac{a}{k}\right) da}}{k \cdot \mu_0} \tag{9}$$

Then do the variable substitution $a=ka_0$ to the integral on the right side of Equation (9):

$$\begin{aligned} \frac{\sigma_a}{\mu_a} &= \frac{\sqrt{\int_{-\infty}^{+\infty} (a - k\cdot\mu_0)^2 \cdot \frac{1}{k} p_0\left(\frac{a}{k}\right) da}}{k \cdot \mu_0} \\ &= \frac{\sqrt{\int_{-\infty}^{+\infty} (ka_0 - k\mu_0)^2 \cdot \frac{1}{k} p_0(a_0) d(ka_0)}}{k\mu_0} \\ &= \frac{\sqrt{\int_{-\infty}^{+\infty} k^2 (a_0 - \mu_0)^2 \cdot \frac{1}{k} p_0(a_0) \cdot k da_0}}{k\mu_0} \\ &= \frac{\sqrt{k^2} \cdot \sqrt{\int_{-\infty}^{+\infty} (a_0 - \mu_0)^2 \cdot p_0(a_0) da_0}}{k\mu_0} \end{aligned} \tag{10}$$

Remember that the variables a and a_0 represent the amplitude value, which is non-negative. Therefore, k is also non-negative. Then Equation (10) can be rewritten as:

$$\frac{\sigma_a}{\mu_a} = \frac{\sqrt{\int_{-\infty}^{+\infty} (a_0 - \mu_0)^2 \cdot p_0(a_0) da_0}}{\mu_0} \tag{11}$$

Notice that the numerator of the right side of Equation (11) is just the standard deviation of a_0. Therefore,

$$\frac{\sigma_a}{\mu_a} = \frac{\sigma_0}{\mu_0} \tag{12}$$

Notice that the right side of Equation (12) is constant given the prototype distribution $p_0(a_0)$. Therefore, the standard deviation coefficient of a is consistent whatever the scaling factor k is. Therefore, this amplitude model accords well with the property of "consistent standard deviation coefficient". If

the prototype pdf $p_0(a_0)$ is determined, the pdf of any frequency's amplitude a can then be derived by $a=\mu a_0$, where μ is a's expectation.

C. *The prototype pdf for ω_k's amplitude spectrum value*

As shown in Fig. 6(b) to Fig. 11(b), for each frequency component, a curve of amplitude distribution after expectation-normalization can be obtained. And these curves are very close to each other. The curves in Fig. 6 to Fig. 11 are averaged. The averaged curve is shown as a black one (the central curve mentioned in Section 3.1), surrounding by the belt area formed by those curves estimated for each ω_k.

Based on these results, a probability distribution is proposed for the average curve (the prototype pdf). The Weibull distribution is used as that prototype distribution, which is adaptive to represent multiple distributions (including the exponential distribution, Rayleigh distribution, Gaussian distribution, etc.) by varying the shape parameter of the function [47-50]. Such generalization ability is quite suitable for the study here. There are two other reasons to use Weibull distribution. First, the amplitude spectrum data is non-negative, which suits the requirement of the Weibull distribution. Second, in the experiment all the estimated distribution curves of expectation-normalized amplitude data have single-peak shape (as shown in Fig. 6 to Fig. 11), which also suits the characteristic of Weibull distribution function.

The Weibull distribution is expressed as a two-parameter function [49,50]:

$$p(x) = b \cdot a^{-b} \cdot x^{(b-1)} \cdot e^{-(\frac{x}{a})^b} \tag{13}$$

where a is the scale parameter and b is the shape parameter. The shape parameter b makes the distribution adaptive to represent different distribution types [47-50]. If b=1, Equation (13) reduces to the exponential distribution. If b=2, it turns to the Rayleigh distribution. If b=3, it well approximates the Gaussian distribution. Therefore, it is highly flexible in fitting experimental data.

For unvoiced pronunciation, in order to estimate Weibull parameters a and b for the prototype pdf, all the expectation-normalized amplitude data of every frequency component are used as a whole data set, since all the frequency components share a unique prototype of amplitude distribution. The statistic toolbox in Matlab is used to estimate Weibull parameters a and b. The estimation results are shown in Table 2.

The results in Table 2 indicate that different pronunciations have obviously different shape parameter values of b, but their scale parameters a are similar due to the preprocessing of expectation-normalization. The shape parameter values approximately range in (1.5, 2.1), which indicate the probability distribution falls in between the exponential distribution (with b=1) to the Rayleigh distribution (with b=2), and with an obvious tendency to the Rayleigh distribution. Using the Weibull distribution here has obvious advantage over those distributions of fixed shape, because the study is mainly based on experimental data, and there is little prior knowledge which can determine the distribution of the spectrum data. There is an interesting phenomenon found in the results that the pronunciations of the same phoneme by different speakers or different recording platform result in same parameters (such as the unvoiced [a] or [ə] in Table 2), which may be utilized in whisper voice recognition.

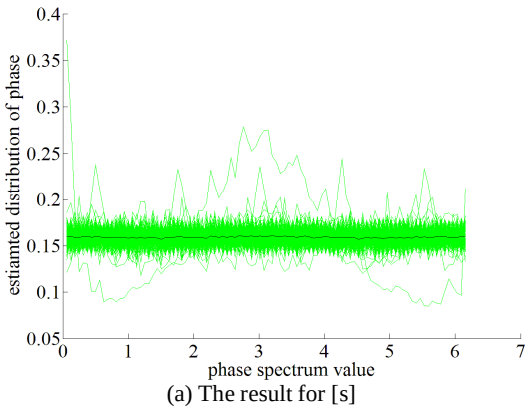
Table 2 The estimated parameters of Weibull distribution for the expectation-normalized amplitude spectrum data

Unvoiced phonation	**scale parameter *a***	**shape parameter *b***
[s] (male)	1.1228	1.6855
[θ] (male)	1.1255	2.0888
[f] (male)	1.1285	2.0261
[h] (male)	1.1234	1.6985
unvoiced [a] (male)	1.1223	1.6758
unvoiced [ə] (male)	1.1154	1.5347
unvoiced [e] (male)	1.1232	1.7170
unvoiced [i] (male)	1.1248	1.7538
unvoiced [a] (female)	1.1223	1.6758
unvoiced [ə] (female)	1.1154	1.5347
unvoiced [a] (male; other recording platform)	1.1223	1.6758
unvoiced [ə] (male; other recording platform)	1.1154	1.5347

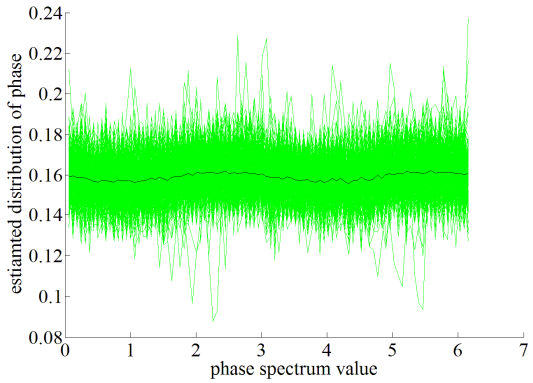
IV. THE MODEL OF THE SHORT-TIME SPECTRUM FOR UNVOICED PRONUNCIATION

A. *The analysis of phase distribution*

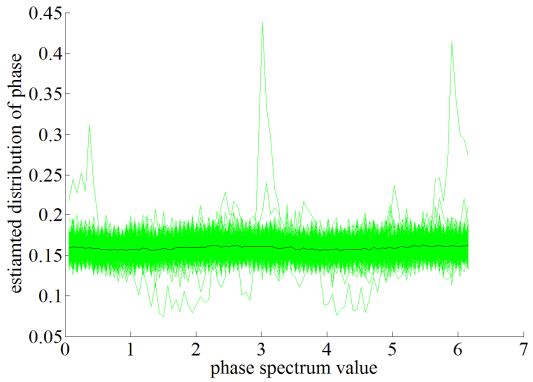
The spectrum value is usually complex number, whose modulus represents the amplitude and the argument represents the phase. Besides the amplitude, the phase distribution is also studied for sustained unvoiced pronunciation. The similar method is used to estimate the phase probability distribution as in Section 3.1, except that the range of phase value is defined as $[-\pi, \pi]$. The phase distribution is estimated for each frequency component respectively. Some of the results are shown in Fig. 12 for the pronunciation of [h], [s], unvoiced [a] and unvoiced [e]. The experimental results indicate a uniform distribution for the phase spectrum value.



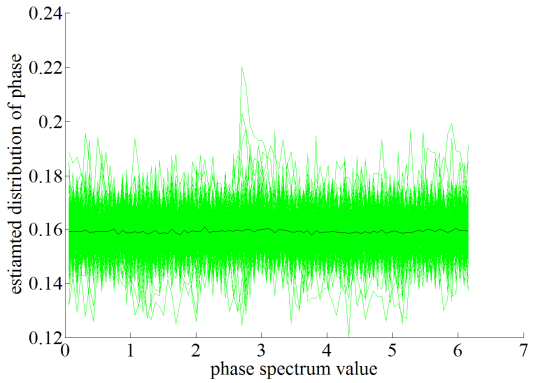
(a) The result for [s]



(b) The result for [h]



(c) The result for unvoiced [a]



(d) The result for unvoiced [e]

Fig. 12. Some results of the estimation of phase distribution

B. The model of short-time amplitude and phase spectrum

By combining the model of amplitude pdf proposed in Section 3 and the phase distribution in Section 4.1, a model is proposed for the short-time spectrum of an unvoiced phoneme:

(1) The short-time spectrum of unvoiced pronunciation is random; for each frequency component ω_k, its amplitude can be modeled as Weibull distribution; its phase can be modeled as uniform distribution in the range of [-π, π].

(2) The standard deviation coefficient of amplitude is consistent for all frequency components.

(3) For a specific unvoiced phoneme, all the frequency components have a common prototype of amplitude distribution a_0. a_0 is the prototype random variable of Weibull distribution. For any frequency component ω_k, the probability distribution of its amplitude is modeled as $\mu_k a_0$, where μ_k is the expectation.

Although the above model is based on the experiments for sustained unvoiced pronunciation, due to the short-time stable property of speech, the above model can also be valid for a short period (such as 16ms) during which the signal is considered as stable.

V. Synthesis of Whispered Pronunciation Based on the Short-time Spectrum Model

In this section, the signal of sustained unvoiced pronunciation is synthesized based on the proposed model of short-time spectrum. The purpose of the synthesis here has two aspects. The first one is to generate synthesized signal with the same perception quality as the original pronunciation. The second one is to experimentally verify the proposed model. Based on the model proposed above, random short-time spectrum data can be artificially generated, and the synthesis of unvoiced pronunciation can be implemented by the reverse STFT.

There are three steps for the proposed synthesis:

Step 1: Estimate the model parameters from the original signal of unvoiced pronunciation. The key parameters of the proposed model are the average amplitude value $\mu(\omega_k)$ for each frequency ω_k, and the two parameters a and b of the Weibull distribution $p_0(x)$ for the expectation-normalized amplitude data.

STFT is performed on the original signal to get the group of short-time spectrums, and the average value of amplitude is estimated for each frequency. Then for each frequency ω_k, its amplitude value for each signal frame is normalized by dividing the corresponding amplitude average $\mu(\omega_k)$, and the two parameters of Weibull distribution are estimated based on the normalized amplitude data of all frequencies as a whole data set.

Step 2: Generate random amplitude and phase values of each ω_k for each synthesized frame. For each frequency component, the amplitude value is generated according to the Weibull distribution, and the phase value is generated according to uniform distribution.

The Weibull distribution for the expectation-normalized amplitude data can be determined by the two parameter s a and b (which are estimated in Step 1):

$$p_0(x) = b \cdot a^{-b} \cdot x^{(b-1)} \cdot e^{-(\frac{x}{a})^b} \tag{14}$$

where $p_0(x)$ is the prototype pdf of amplitude, a and b are the scale parameter and shape parameter respectively. The actual amplitude pdf of the k-th frequency component should be deduced from $p_0(x)$. The actual amplitude data for the k-th frequency component has the expectation value $\mu(\omega_k)$, and the corresponding amplitude distribution can be deduced as:

$$p_k(x) = \frac{1}{\mu(\omega_k)} \cdot p_0\left(\frac{x}{\mu(\omega_k)}\right) \tag{15}$$

where $p_k(x)$ is the amplitude pdf of ω_k. Then the amplitude data for ω_k can be generated artificially according to Equation (15). The phase value can be generated according to the uniform

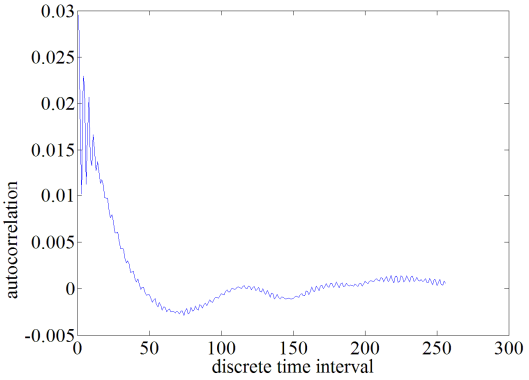
distribution in the range of [-π, π]. Because the DFT results of real signal has the conjugate symmetric property, the short-time spectrum $X_i(k)$ can be constructed with the artificially generated amplitude and phase values in a conjugate symmetric way. By Step 2, a group of artificial spectrum data of signal frames can be generated for the corresponding whisper pronunciation.

Step 3: Time domain signal construction by reverse STFT. The reverse STFT transforms a group of successive short-time spectrum to the time domain signal frames. IDFT (inverse discrete Fourier transform) is performed on each short-time spectrum:

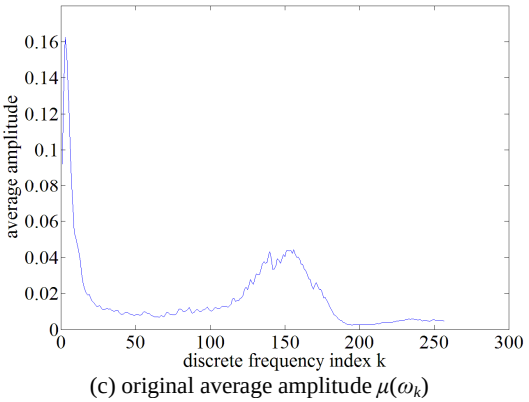
$$x_i(n) = \frac{1}{N}\sum_{k=0}^{N-1} X_i(k)\cdot e^{j\frac{2\pi}{N}kn} \tag{16}$$

where $x_i(n)$ is the i-th synthesized frame, and $X_i(k)$ is its corresponding spectrum. Then the successive frames are combined to produce the synthesized signal by an overlap and adding process: two adjacent frames are overlapped by half of the frame length, and then added. The final result is the synthesized signal.

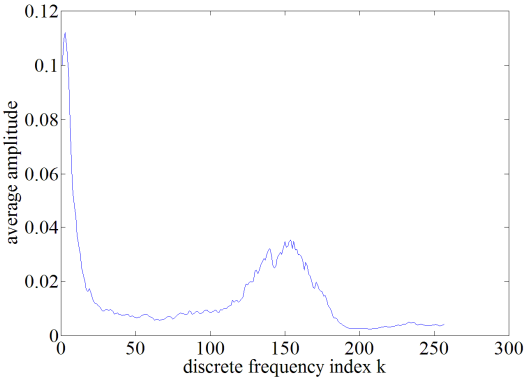
In the listening test, the listeners (ten male and ten female listeners of the age 19-25 with normal audition), the synthesized signals have identical quality of auditory perception compared to the corresponding original whisper pronunciation. On the other hand, the average amplitude $\mu(\omega_k)$ and the signal autocorrelation of the original and synthesized pronunciation are computed for comparison. Some of the results are shown in Fig. 13 to Fig. 16. It is indicated that the time-domain autocorrelation of the synthesized signals are linear proportional to those of the original pronunciations. Because the power spectrum of a stochastic signal can be determined by its time-domain autocorrelation, similar autocorrelation functions correspond to similar power spectrum. That is why the synthesized signals have identical perception quality compared to the original ones. This is also indicated by the similar average amplitude curve $\mu(\omega_k)$ of the original and the synthesized signal, which is shown in Fig. 13 to Fig. 16. The high quality of synthesis for unvoiced pronunciation proves the effectiveness of the model proposed in Section 4.2.



(b) autocorrelation of the synthesized signal

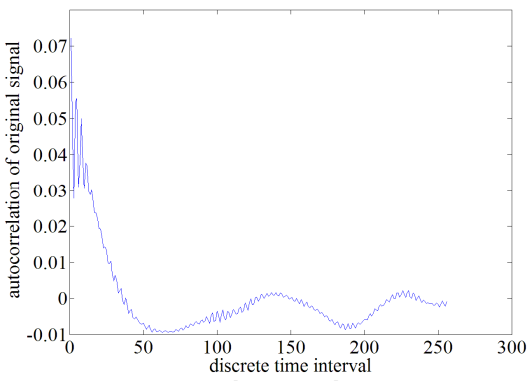


(c) original average amplitude $\mu(\omega_k)$

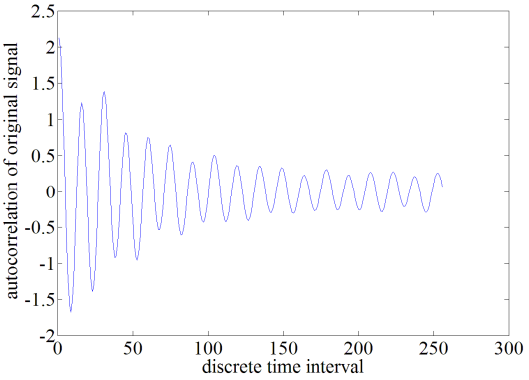


(d) $\mu(\omega_k)$ of the synthesized signal

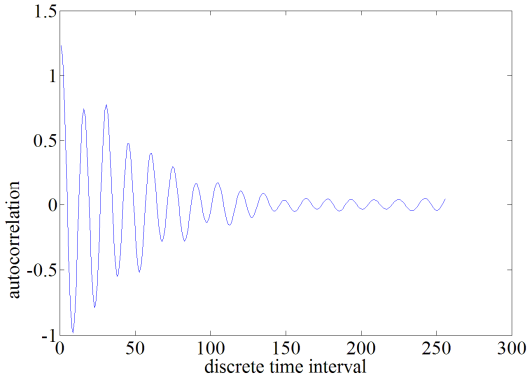
Fig. 13. The comparison between the original unvoiced signal and the synthesized signal for [θ]



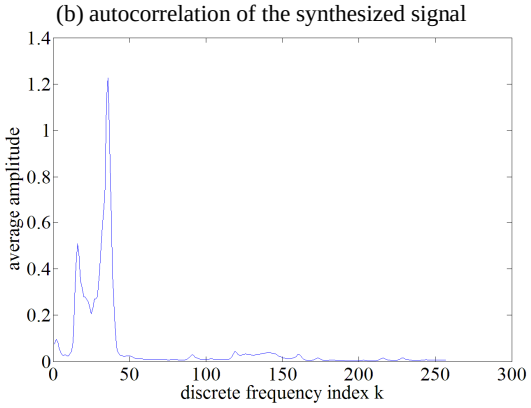
(a) original autocorrelation



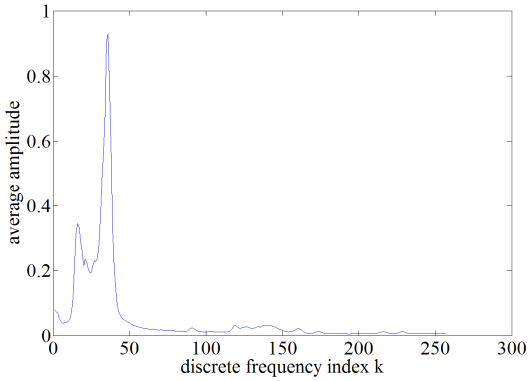
(a) original autocorrelation



(b) autocorrelation of the synthesized signal

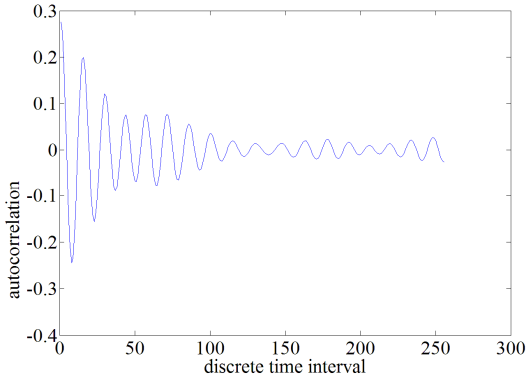


(c) original average amplitude $\mu(\omega_k)$

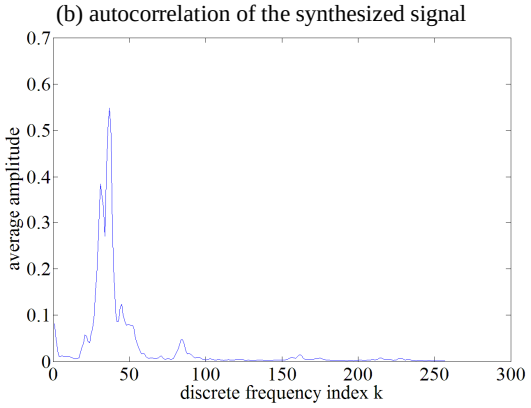


(d) $\mu(\omega_k)$ of the synthesized signal

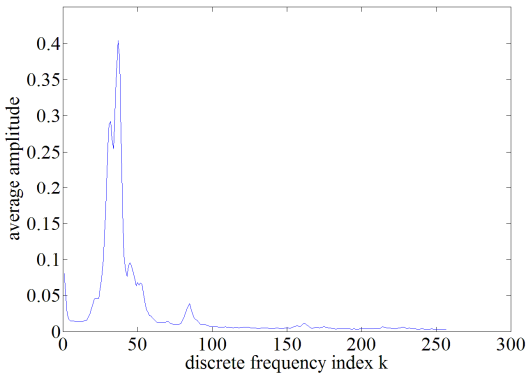
Fig. 14. The comparison between the original unvoiced signal and the synthesized signal for [h]



(b) autocorrelation of the synthesized signal

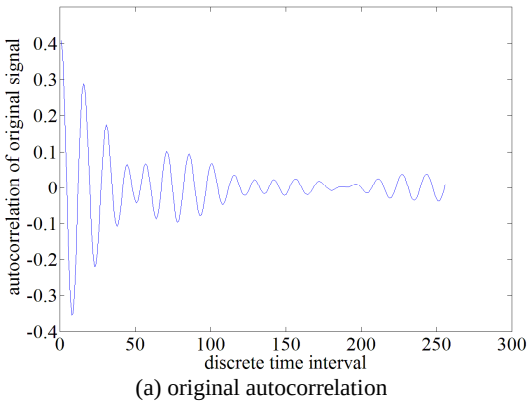


(c) original average amplitude $\mu(\omega_k)$

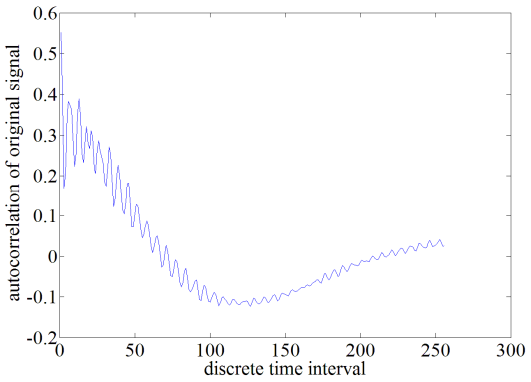


(d) $\mu(\omega_k)$ of the synthesized signal

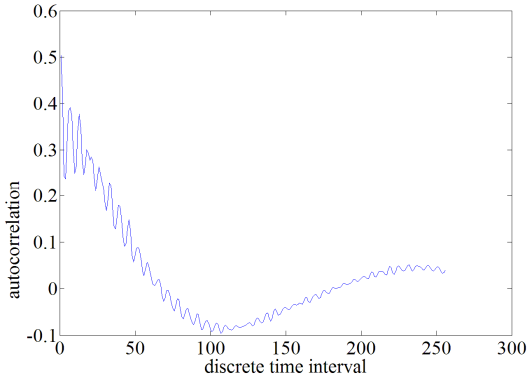
Fig. 15. The comparison between the original unvoiced signal and the synthesized signal for unvoiced [a]



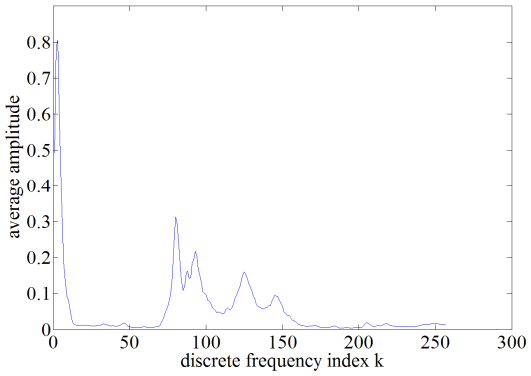
(a) original autocorrelation



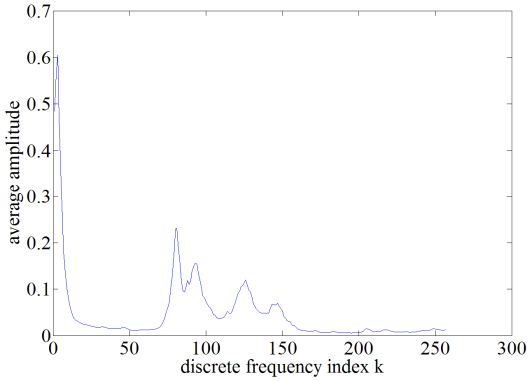
(a) original autocorrelation



(b) autocorrelation of the synthesized signal



(c) original average amplitude $\mu(\omega_k)$



(d) $\mu(\omega_k)$ of the synthesized signal

Fig. 16. The comparison between the original unvoiced signal and the synthesized signal for unvoiced [i]

VI. Conclusion and Discussion

In this paper, we study the signal features of unvoiced pronunciation in whisper speech. The stochastic feature of the short-time spectrum for unvoiced pronunciation is investigated. For the short-time amplitude spectrum of single unvoiced phonemes, the relationship between the amplitude's expectation and standard deviation is analyzed for individual frequency components. In the experiments, for an unvoiced phoneme, the ratio of standard deviation to expectation (also called the standard deviation coefficient) is proved to be consistent for all the frequency components. This new feature is related to the physical aero-acoustic process of whisper pronunciation. Based on this new feature, a stochastic model of amplitude spectrum value is proposed, in which all the frequency components share a common prototype of pdf. Moreover, the probability distribution of amplitude after expectation-normalization is estimated. The probability distribution of phase is also studied.

By combining the proposed amplitude and phase distribution, a stochastic model of short-time spectrum is proposed for whisper pronunciation, with a Weibull distribution for the expectation-normalized amplitude, and a uniform distribution for the phase. Based on this model, an efficient synthesis method is presented whisper speech, which yields equivalent quality of auditory perception as the original unvoiced pronunciation, and also similar autocorrelation compared to that of the original signal. The effectiveness of the synthesis proves the validity of the proposed stochastic model of unvoiced pronunciation in frequency domain.

The work in this paper indicates that, besides the general statistic properties of speech signals in daily communication, it is worthwhile to study the stochastic properties of specific pronunciation types, which is on a different level of randomness for speech signal. Just as the whisper pronunciation studied in this paper, specific type of pronunciation signal has more detailed and unique features compared to normal speech signals. Since the whisper pronunciation is physically based on an aerodynamic process, the signal feature revealed in this paper is also inspiring for the research on acoustic signal produced by physical aerodynamic process.

References

[1] M. Cotescu, T. Drugman, G. Huybrechts, J. Lorenzo-Trueba and A. Moinet, 2020, Voice Conversion for Whispered Speech Synthesis, IEEE Signal Processing Letters, vol. 27, 186-190.

[2] A. R. Naini, A. R. M. V. and P. K. Ghosh, 2019, Formant-gaps Features for Speaker Verification Using Whispered Speech, 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 6231-6235.

[3] [3] V. Vestman, D. Gowda, Md Sahidullah, P. Alku, T. Kinnunen, 2018, Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction, Speech Communication, Vol 99, 62-79.

[4] F. Kelly and J. H. L. Hansen, 2018, Detection and Calibration of Whisper for Speaker Recognition, 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 1060-1065.

[5] D. T. Grozdic, S. T. Jovicic, 2017, Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 12, 2313-2322

[6] H. Konno, M. Kudo, H. Imai, M. Sugimoto, 2016, Whisper to normal speech conversion using pitch estimated from spectrum, Speech Communication, Vol. 83, 10-20.

[7] H. R. Sharifzadeh, I. V. McLoughlin, M. J. Russell, 2012, A Comprehensive Vowel Space for Whispered Speech, Journal of Voice, Vol 26, Issue 2, e49-e56.

[8] J. Sundberg, R. Scherer, M. Hess, F. Muller, 2010, Whispering-A Single-Subject Study of Glottal Configuration and Aerodynamics, Journal of Voice, Vol 24, Issue 5, 574-584.
[9] S. T. Jovicic, Z. Saric, 2008, Acoustic Analysis of Consonants in Whispered Speech, Journal of Voice, Volume 22, Issue 3, 263-274.
[10] M. Parmar, S. Doshi, N. J. Shah, M. Patel and H. A. Patil, 2019, Effectiveness of Cross-Domain Architectures for Whisper-to-Normal Speech Conversion, 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 1-5.
[11] D. Sivan and C. Gopakumar, 2017, Emotion recognition and spoof detection from whispered speech, 2017 International Conference on Computing Methodologies and Communication (ICCMC), Erode, 1091-1095.
[12] G. Srinivasan, A. Illa and P. K. Ghosh, 2019, A Study on Robustness of Articulatory Features for Automatic Speech Recognition of Neutral and Whispered Speech, 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 5936-5940.
[13] O. Perrotin and I. V. McLoughlin, 2020, Glottal Flow Synthesis for Whisper-to-Speech Conversion, IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, 889-900.
[14] K. Khoria, M. R. Kamble and H. A. Patil, 2020, Teager Energy Cepstral Coefficients for Classification of Normal vs. Whisper Speech, 28th European Signal Processing Conference (EUSIPCO), Amsterdam, 1-5.
[15] R. Konnai, R. C. Scherer, A. Peplinski, K. Ryan, 2017, Whisper and Phonation: Aerodynamic Comparisons Across Adduction and Loudness, Journal of Voice, Vol 31, Issue 6, 773.e11-773.e20.
[16] Y. Okada et al., 2020, Effects of Touch Behaviors and Whispering Voices in Robot-Robot Interaction for Information Providing Tasks, 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), Naples, Italy, 7-13.
[17] S. Petridis, J. Shen, D. Cetin and M. Pantic, 2018, Visual-Only Recognition of Normal, Whispered and Silent Speech, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, 6219-6223
[18] Y. Zhao, W. Lin, 2016,Study of the formant and duration in Chinese whispered vowel speech, Applied Acoustics, Vol. 114, 240-243.
[19] T. G. Csapo, G. Nemeth, M. Cernak and P. N. Garner, 2016, Modeling unvoiced sounds in statistical parametric speech synthesis with a continuous vocoder, 2016 24th European Signal Processing Conference (EUSIPCO), Budapest, 1338-1342.
[20] Garofolo J., Lamel L., Fisher W., Fiscus J., Pallett D., Dahlgren N., Zue V., 1993. TIMIT Acoustic-phonetic continuous speech corpus, Linguistic Data Consortium, Philadelphia.
[21] Hirsch, H.-G., Pearce, D., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. Automatic Speech Recognition: Challenges for the Next Millennium (ISCA ITRW ASR2000), Paris, France, pp. 181-188.
[22] Gazor S., Zhang W., 2003. Speech probability distribution, IEEE Signal Processing Letters, 10(7), 204-207.
[23] Davenport W. B., 1952. An experimental study of speech wave probability distributions. J. Acoust. Soc. Amer., 24(4), 390–399.
[24] Richards D. L., 1964. Statistical properties of speech signals, Proc. Inst. Elect. Eng., 111(5), 941–949.
[25] Paez M. D., Glisson T. H., 1972. Minimum mean-square error quantization in speech. IEEE Trans. Comm., 20, 225–230.
[26] Jayant N. S., Noll P., 1984. Digital coding of waveforms. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.
[27] Shin J. W., Chang J.-H., Kim N. S., 2005. Statistical modeling of speech signals based on generalized gamma distribution. IEEE Signal Processing Letters, 12(3), 258-261.
[28] Jensen J., Batina I., Hendriks R. C., Heusdens R., 2005. A study of the distribution of time-domain speech samples and discrete Fourier coefficients. Proceedings of SPS-DARTS (The first annual IEEE BENELUX/DSP Valley Signal Processing Symposium), pp. 155-158.
[29] Tashev I., Acero A., 2010. Statistical modeling of the speech signal, International Workshop on Acoustic, Echo, and Noise Control (IWAENC), Tel Aviv, Israel.
[30] Erkelens J. S., Jensen J., Heusdens R., 2007. Speech enhancement based on Rayleigh mixture modeling of speech spectral amplitude distributions. 15th European Signal Processing Conference (EUSIPCO 2007), pp. 65-69.
[31] Martin R., 2005. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. IEEE Transactions on Speech and Audio Processing. 13(5), 845-856.
[32] Loizou P. C., 2007. Speech enhancement, theory and practice. Taylor & Francis, New York, NY, USA, 1st edition.
[33] Boubakir C., Berkani D., 2010. Speech enhancement using minimum mean-square error amplitude estimators under normal and generalized gamma distribution. Journal of Computer Science, 6(7), 700-705.
[34] Borgstrom B. J., Alwan A., 2011. Log-spectral amplitude estimation with Generalized Gamma distributions for speech enhancement. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4756-4759.
[35] Sohn J., Kim, N. S., Sung W., 1999. A statistical model-based voice activity detection. IEEE Signal Processing Letters, 6, 1-3.
[36] Rabiner L., Juang B.-H., 1993. Fundamentals of speech recognition. Prentice-Hall International, Inc..
[37] Huang J. and Zhao Y., 2000. A DCT-based fast signal subspace technique for robust speech recognition. IEEE Trans. Speech Audio Processing, 8, 747–751.
[38] Teixeira J. P., Oliveira C., Lopes C., 2013. Vocal acoustic analysis - jitter, shimmer and HNR parameters. Procedia Technology, 9, 1112-1122.
[39] Ghosh P. K., Narayanan S. S., 2011. Joint source-filter optimization for robust glottal source estimation in the

presence of shimmer and jitter. Speech Communication, 53(1), pp. 98-109.
[40] Farrus M., Hernando J., 2009. Using jitter and shimmer in speaker verification. IET Signal Processing, 3(4), pp. 247-257.
[41] Sinder D. J., Krane M. H., Flanagan J. L.,1998. Synthesis of fricative sounds using an aeroacoustic noise generation model. Proceedings of 16th International Congress Acoustics, 1, pp. 249–250.
[42] Mittal R., Erath B. D., Plesniak M. W., 2013. Fluid dynamics of human phonation and speech. Annual Review of Fluid Mechanics, 45, 437-467.
[43] Lu X. B., Thorpe C. W., Cater J. E., Hunter P. J., 2011. Aeroacoustic modeling of frictives /s/ and /sh/. Proceedings of the 18th international congress on sound & vibration, pp. 373-380.
[44] Sinder D., Richard G., Duncan H., Lin Q., Flanagan J., 1996. A fluid flow approach to speech generation, First ETRW on Speech Production Modelling, pp. 203-206.
[45] Hirschberg A., 1992. Some fluid dynamic aspects of speech. Bulletin de la Communication Parlee, 2, 7-30.
[46] McGowan R. S., 1987. An aeroacoustics approach to phonation: some experimental and theoretical observations. Haskins Laboratories: Status Report on Speech Research SR-86/87, pp. 107-116.
[47] Weibull W., 1951. A statistical distribution function of wide applicability. Journal of Applied Mechanics, 18, 293-297.
[48] Lindquist E. S., 1994. Strength of materials and the Weibull distribution, Probabilistic Engineering Mechanics, 9(3), 191-194.
[49] Khaledi B.-E., Kochar S., 2006. Weibull distribution: Some stochastic comparisons results. Journal of Statistical Planning and Inference, 136(9), 3121-3129.
[50] Szymkowiak M., Iwinska M., 2016. Characterizations of Discrete Weibull related distributions. Statistics & Probability Letters, 111, 41-48.