A New Method for the Objective Perceptual Measurement of Chinese Initials

Sai Chen[1] , Hongcui Wang[1,*] , Jia Jia[2] and Jianwu Dang[1,3]
[1] Tianjin Key Laboratory of Cognitive Computing and Application, School of Computer Science and Technology, Tianjin University, China
E-mail: chensai@tju.edu.cn; hcwang@tju.edu.cn
[2] Department of Computer Science and Technology, Tsinghua University, China
E-mail: jjia@tsinghua.edu.cn
[3] School of Information Science, Japan Advanced Institute of Science and Technology, Japan
E-mail: jdang@jaist.ac.jp

Abstract— **Many works have been done in the methods of the perception measurements for speech sound. However, most of them are subjective measurement alone for perception aspects. In this paper, we try to give a new method of objective perception measurement for Chinese initials and investigate the relationship between the acoustic features and the perception measurement. To do so, we discuss which acoustic features and their combinations are the most consistent with the real perception of Chinese initials. We propose a method to obtain an objective perception measure based on the acoustic features, where the acoustic distance has a monotonic relation with the perceptual distance for Chinese initials. The Spearman's rank correlation coefficient is enhanced from 0.6328 to 0.6498 by adding the time-domain features to the feature vector of each initial. Finally, we propose a new formula to measure the perceptual distance between different types of initials objectively by using the chosen acoustic features.**

I. INTRODUCTION

From phonetic view of point, the place and manner of the articulation are applied to classify Chinese initials. And the statistical and psychological methods are used to explore the perceptual characteristics [1]. The characteristics of phonation and articulation, such as voiced or voiceless, aspirated or unaspirated, and fricative or frictionless, are the most important factors that influences the perception of initials [2]. A perceptual measurement based on LPC has been proposed for Chinese finals in [3], which makes it easier to evaluate the equivalence of different audiometric word lists. The acoustic features most commonly used are Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) features [4]. Both MFCC and PLP are tested with and without 'pitch' information using the same back-end on an English consonants corpus and the results are compared with human listener results at the level of articulatory feature classification, which shows that no representation reaches the levels of human performance but PLP has higher accuracies for most manner values on English consonants than MFCC [5]. However, the perception of Chinese initials, which are not exactly the same as English consonants, is more difficult for humans, especially for patients, than that of Chinese finals. Hence, it is very important to do the research on the perceptual characteristics of Chinese initials. A method has been proposed in [6], where six PLP coefficients (the 4th, 5th, 8th, 9th, 10th and 11th order of the 12 coefficients) with the Shortest and Chebyshev as the inter-cluster and intra-cluster distance measures are used to calculate the acoustic distance. Then the perceptual distance is obtained from a speech audiometry. The Spearman's rho of the two types of distance is finally calculated, which is 0.6328.

In this paper, we systematically test the time-domain features of Chinese initials by carrying out two experiments with respect to acoustic space and perceptual space, respectively. The experimental results show that the Spearman's rho is 0.4678, which is smaller than that using PLP. However, after we combine the PLP and the time-domain features, we obtain a maximum Spearman's rho of 0.6498, which is larger than the original method that uses PLP. Finally, we propose a new formula to measure the perceptual distance between two types of initials objectively based on the chosen acoustic features.

II. ACOUSTIC EXPERIMENT AND RESULTS

A. Acoustic Distance

All the phonemes of Chinese initials are divided into 21 categories. However, they are not identical when joined with different finals. Because clustering is adaptable to changes and helps single out useful features that distinguish different group and it can be used as a standalone tool to gain insight into the distribution of data [7], we consider each category as a cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. We then define the acoustic distance between two

*The corresponding author

categories of initials as the distance between two clusters. We use hierarchical methods to analyze the relationship between different kinds of initials, since it leads to smaller computation costs by not having to worry about a combinatorial number of different choices [8], which is suitable for the task attempting to use as many dissimilarity measures as possible.

B. Dissimilarity Metrics

Before we calculate the distance between clusters, we should single out the dissimilarity metric between samples of various initials, which is the key component in clustering analysis. The Euclidean distance between two samples of initials was used in [2]. In this paper, we use up to 10 types of dissimilarity measures of objects (including 4 variations of Minkowskia, i.e. the exponent is equal to 3, 4, 5, and 10, respectively) [11], which are listed in table 1.

In order to calculate the acoustic distance between two types of initials, we also need to choose the best distance measures between clusters. Seven widely used measures for distance between clusters [12] are used in this paper. They are listed in table 2.

C. Data Corpus

The speech material is a standard corpus provided from the General Hospital of the People's Liberation Army (PLAGH) in speech audiometry, which are recorded in an acoustically isolated booth by a male broadcaster. The frequency of sampling is 44100 Hz. There are 470 Chinese monosyllables in the corpus and they consist of all the categories of initials (/b/, /c/, /ch/, /d/, /f/, /g/, /h/, /j/, /k/, /l/, /m/, /n/, /p/, /q/, /r/, /s/, /sh/, /t/, /x/, /z/, /zh/) excluding zero initials (/y/ and /w/), and almost all possible combinations of initials, finals and tones. Each monosyllable is segmented into two parts, the initial and final, and labeled manually using the software called VisualSpeech developed by Tsinghua University.

TABLE I
DISSIMILARITY MEASURES

NAME	FORMULA	
MANHATTAN	$D(P,Q)=\sum_{i=1}^{n}(p_i-q_i)$	(1)
EUCLIDEAN	$D(P,Q)=[\sum_{i=1}^{n}(p_i-q_i)^2]^{\frac{1}{2}}$	(2)
STANDARDIZED EUCLIDEAN	$D(P,Q)=[\sum_{i=1}^{n}\left(\frac{p_i-q_i}{s_k}\right)^2]^{\frac{1}{2}}$	(3)
CHEBYSHEV	$D(P,Q)=\lim_{k\to\infty}[\sum_{i=1}^{n}(p_i-q_i)^k]^{\frac{1}{k}}$	(4)
COSINE	$D(P,Q)=1-\cos<\boldsymbol{P},\boldsymbol{Q}>$	(5)
CORRELATION	$D(P,Q)=1-\rho_{\boldsymbol{PQ}}$	(6)
MINKOWSKIA	$D(P,Q)=[\sum_{i=1}^{n}\lvert p_i-q_i\rvert^t]^{\frac{1}{t}}$	(7)

TABLE II
DISTANCE MESSURES BETWEEN CLUSTERS

Name	Definition
Furthest	The longest distance between two points in each cluster.
Shortest	The shortest distance between two points in each cluster.
UPGMA	The average of all distances between pairs of objects, i.e. the mean distance between elements of each cluster.
WPGMA	The weighted average distance between two samples in the two clusters respectively.
UPGMC	The Euclidean distance between their centroids.
WPGMC	The Euclidean distance between their weighted centroids.
Ward	The distance between two clusters is how much the sum of squares will increase when we merge them.

D. Clustering Analysis and Results

We extract 6 types of time-domain features of each frame, the duration, the short-time zero crossing rate, the short-time average energy, the ratio of maximum short-time average energy to the average energy, the ratio of minimum short-time average energy to the average energy, and the ratio of the short-time zero crossing to the short-time average energy. Then we calculate the mean of each type of features for all the frames in an initial to form the feature vector. Hence, the feature vector of each initial consists of six elements.

Normalization is particularly useful for distance measurements such as clustering, which gives all attributes an equal weight. Here, the features are normalized using a variation of the z-score normalization:

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A} \tag{8}$$

where $\bar{A}$ and σ_A are the mean and standard deviation of each of the 6 features respectively.

We generate all possible combinations of the time-domain features (the total number of possible combinations of features is up to 63). We define the accuracy of hierarchical clustering of the initials, Acc, as follows:

$$\mathrm{Acc} = \begin{cases} \frac{n_i}{N_i}, & \frac{n_i}{N_i} \geq 0.6 \\ 0, & \frac{n_i}{N_i} < 0.6 \end{cases} \tag{9}$$

where n_i is the number of the samples of the i^{th} category of initial which are grouped into a cluster, and N_i is the number of the samples of the i^{th} category of initial.

We calculate Acc using all possible combinations of the time-domain features and all dissimilarity metrics (43 in total), and then calculate the average accuracy of hierarchical clustering of initials, $\overline{\mathrm{Acc}}$, which is defined as the arithmetic mean of 21 Acc corresponding to the 21 categories of initial. We expect $\overline{\mathrm{Acc}}$ to be as large as possible. We also calculate the variance for each distance metric. The results, where the arithmetic means are larger than 0.7, are listed in table 3.

The experimental results show that the clustering using Shortest and Manhattan has the highest $\overline{\mathrm{Acc}}$. Hence, we can infer that the Shortest and Manhattan are the most compatible distance metric as the inter-cluster and intra-cluster metrics, respectively, with the time-domain features.

III. PERCEPTUAL EXPERIMENT AND RESULTS

The perceptual distance between two types of initials is defined as follows:

$$P_{uv} = 1 - \frac{Pr\{I_u \text{ is misheard as } I_v\} + Pr\{I_v \text{ is misheard as } I_u\}}{2} \quad (10)$$

where I_u and I_v are the two types of initials, and the probability of mishearing is calculated by the confusion matrix obtained in the speech audiometry. The perceptual distance reflects how far one initial from another in perceptual space. The more confusing the two types of initials are, the smaller the perceptual distance between them is.

We design an experiment to get the perceptual distance between each pair of initials. Twenty subjects at the age of about 25 without hearing loss or ear diseases taking part in the experiment. First, we set an initial sound intensity for each subject and pick up five monosyllables randomly from the corpus to present to the subject. Then, the subject is asked to answer which initial it is. When the five monosyllables are all played, we compare the answers given by the subject to the right answers to calculate the recognition probability. If the accuracy is higher than 50%, we decrease the sound intensity; otherwise, we increase the sound intensity. Finally, we get the Speech Reception Threshold (SRT), which is the sound intensity at which the subject gains 50% recognition probability [15]. We then generate a random permutation of all the 470 monosyllables in the corpus and play them to each subject with the sound intensity of SRT. Based on the answers given by each subject, a 21-by-21 matrix is constructed, where the element $e(i, j)$ indicates the count of the ith initial misheard as the jth initial. However, while in experiment, subjects may mishear some initials not because the initials are easily confused, but because the subjects themselves are absent-minded, weary or affected by the equipment. The small probability events, caused by different subjects or equipment, is reflected in the confusion matrix as elements with very small values. We eliminate those errors by setting the elements less than or equal to 0.01 in the confusion matrix to be zero. Finally, we use (10) to transform the confusion matrix into perceptual distance matrix, a 21-by-21 matrix, where the element $p(i, j)$ indicates the perceptual distance between the ith initial and the jth initial.

IV. The Relationship between Acoustic Distance and Perceptual Distance

We validate the feature extraction method and two distance measures using a nonparametric measure, called Spearman's rank correlation coefficient (or Spearman's rho). One of the two variables used in Spearman's rho indicates the perceptual distance (i.e. the elements in perceptual matrix), and the other indicates the acoustic distance (i.e. the elements in the same line and column as those in perceptual matrix). A perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other. It doesn't rely on the assumption that the data are drawn from a given probability distribution, and its interpretation doesn't depend on the population fitting any parametric distributions. Moreover, it's no matter whether the sample size is large or small. These properties are quite useful for our target. We convert acoustic distances and perceptual distances into ranks a_i and p_i, respectively, where identical values are assigned a rank equal to the average of their positions in the ascending order of the values, and the Spearman's rho, ρ, is computed as follows:

$$\rho = \frac{\sum_i (a_i - \bar{a})(p_i - \bar{p})}{\sqrt{\sum_i (a_i - \bar{a})^2 \sum_i (p_i - \bar{p})^2}} \quad (11)$$

We calculate the Spearman's rho using each of the 63 acoustic distance matrices and the perceptual distance matrix. Each acoustic distance is calculated by using the Shortest and Manhattan chosen in Section II D. We then single out the maximum of the 63 Spearman's rho and which time-domain features corresponds to that maximum. The results are listed in Table 4. The symbol j^{th} indicate the j^{th} type of features mentioned in Section II D.

Table IV shows that a Spearman correlation of 0.4678 occurs when all the six time-domain features are selected to compute the acoustic distance. It is smaller than that using PLP in [6]. In order to enhance the Spearman's rho, we combine the time-domain features and PLP together. We extract 12 coefficients of PLP for each frame of an initial, and calculate the mean of coefficients for all the frames as the 12 coefficients of the initial. The PLP coefficients were calculated using the rastamat Matlab toolbox [13] with parameters that resemble feature extraction from the HTK software [14]. According to [6], we use the 4^{th}, 5^{th}, 8^{th}, 9^{th}, 10^{th}, 11^{th} order of the 12 coefficients of each initial, and Shortest and Chebyshev as the inter-cluster and intra-cluster distance measures respectively. We generate all combinations of the six types of time-domain features mentioned in Section II D, each of which is added to the six PLP coefficients to form a new feature vector of the initial. The number of the elements in each vector ranges from 7 to 12. Then, the 63 acoustic distance matrices are computed, each of which corresponds to a combination of the six types of time-domain features. The 63 Spearman's rho

TABLE III
AVERAGE ACCURACY OF HIERARCHICAL CLUSTERING USING TIME-DOMAIN FEATURES

Distance Metrics	$\overline{ACC}$	Var
Shortest-Manhattan	0.8749	0.0008
Shortest-Euclidean	0.8649	0.0008
Shortest-Chebyshev	0.8642	0.0008
Shortest-Minkowskia(exp=10)	0.8635	0.0008
Shortest-Minkowskia(exp=5)	0.8625	0.0008
Shortest-Minkowskia(exp=3)	0.8611	0.0009
Shortest-Minkowskia(exp=4)	0.8606	0.0009
Shortest-Std. Euclidean	0.8533	0.0008
Shortest-Cosine	0.7589	0.0153

TABLE IV
SPEARMAN'S RHO USING TIME-DOMAIN FEATURES

Time-domain Features	Spearman's rho
1^{th}, 2^{th}, 3^{th}, 4^{th}, 5^{th}, 6^{th}	0.4678
1^{th}, 2^{th}, 3^{th}, 4^{th}, 6^{th}	0.4637
1^{th}, 2^{th}, 3^{th}, 4^{th}, 5^{th}	0.4620
1^{th}, 2^{th}, 3^{th}, 5^{th}, 6^{th}	0.4613
1^{th}, 2^{th}, 6^{th}	0.4596
1^{th}, 2^{th}, 3^{th}, 6^{th}	0.4579

corresponding to the 63 acoustic distance matrices are computed later. The Spearman's rho larger than 0.6 are listed in Table 5, with the corresponding time-domain features.

The results show that a Spearman's rho of 0.6498 occurs when the feature vector of each initial consists of the 4th, 5th, 8th, 9th, 10th, 11th order of the 12 PLP coefficients and the initial duration. The two most top-ranking rho, 0.6498 and 0.6441, are both larger than that using PLP proposed in [6]. It shows that the duration of an initial is very important in perceptual measurement of Chinese initials.

Based on the chosen acoustic features mentioned above, we propose a new formula to measure the perceptual distance between different types of initials objectively:

$$D_{IJ} = \min_{\substack{i\in I \\ j\in J}}\{\lim_{n\to\infty}[\sum_{k=0,4,5,\atop 8,9,10,11}\left(s_{ik}-s_{jk}\right)^n]^{\frac{1}{n}} \qquad (12)$$

where D_{IJ} is the perceptual distance between the I^{th} and J^{th} types of initials, s_{ik} and s_{jk} indicate the k^{th} order of the 12 PLP coefficients of the samples belonging to the I^{th} and J^{th} types of initials, respectively, when k is not equal to zero, and s_{i0}, s_{j0}indicate the duration of the samples belonging to the I^{th} and J^{th} types of initials, respectively.

V. CONCLUSIONS

In this paper, we systematically test the time-domain features of Chinese initials by carrying out two experiments with respect to acoustic space and perceptual space, respectively. The experimental results show that the Spearman's rho is 0.4678, which is smaller than that using PLP. However, when we combine the PLP and time-domain features, we obtain a Spearman's rho of 0.6498, which is larger than the method using PLP alone. It shows that the duration of an initial is one important feature in perceptual measurement of Chinese initials. Finally, we propose a new formula to measure the perceptual distance between two types of initials objectively based on the chosen acoustic features.

TABLE V
SPEARMAN'S RHO USING BOTH PLP AND TIME-DOMAIN FEATURES

Time-domain Features	Spearman's rho
1^{th}	0.6498
1^{th}, 4^{th}	0.6441
4^{th}	0.6280
1^{th}, 5^{th}	0.6107
1^{th}, 4^{th}, 5^{th}	0.6061
1^{th}, 6^{th}	0.6050

ACKNOWLEDGMENT

This work is supported in part by the National Basic Research Program of China (No. 2013CB329301), and in part by the national natural science foundation of China under contract No. 61233009, No. 61303109, No. 6117501, No. 61370023 and No. 61003094.

REFERENCES

[1] J. Zhang, S. Qi, and S. Lv. "An Study on Perceptual Structure of Chinese Initials." Acta Psychologica Sinica 1 (1981): 76-85.
[2] J. Jia, Y. Wang, Y. Zhang, et al.. "An Investigation on Calculating Intelligibility Among Chinese Initials." PCC2012
[3] G. Huang, J. Jia, and L. Cai. "A Study on Perceptual Metric Among Chinese Finals Based on LPC." PCC2010
[4] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," Journal Acoust. Soc. Amer., vol. 87, no. 4, pp. 1738–1752, 1990.
[5] Scharenborg, Odette, and M. P. Cooke. "Comparing human and machine recognition performance on a VCV corpus." Proc. Workshop on Speech Analysis and Processing for Knowledge Discovery. 2008
[6] S. Chen, H. Wang, J. Jia, et al. "Comparison of Mel Frequency Ceptrum Coefficient and Perceptual Linear Predictive in Perceptual Measurement of Chinese Initials," unpublished.
[7] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2006.
[8] Johnson, Stephen C. "Hierarchical clustering schemes." Psychometrika 32.3 (1967): 241-254.
[9] E. Schukat-Talamazzini, Automatische Spracherkennung–Grundlagen, statistische Modelle und effiziente Algorithmen.Braunschweig: Vieweg, 1995.
[10] Hönig, Florian, et al. "Revising perceptual linear prediction (PLP)." Proceedings of INTERSPEECH. 2005.
[11] Deza, Michel Marie, and Elena Deza. Encyclopedia of distances. Springer Berlin Heidelberg, 2009.
[12] Murtagh, Fionn. "Complexities of hierarchic clustering algorithms: State of the art." Computational Statistics Quarterly 1.2 (1984): 101-113.
[13] D. P. W. Ellis, "PLP and RASTA (and MFCC, and inversion) in Matlab," 2005, online web resource. [Online]. Available: http://www.ee.columbia.edu/dpwe/resources/matlab/rastamat/
[14] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.2)," Cambridge University, Eng. Dept., 2002, techn. Report.
[15] Boothroyd, Arthur. "The performance/intensity function: an underused resource." Ear and hearing 29.4 (2008): 479-491.