Applied Mechanics and Materials Vols. 411-414 (2013) pp 291-297

doi:10.4028/www.scientific.net/AMM.411-414.291

Comparison of Mel Frequency Ceptrum Coefficient and Perceptual Linear Predictive in Perceptual Measurement of Chinese Initials

Sai Chen[1, a], Hongcui Wang[1, b, *], Jia Jia[2, c], Yeteng An[1, d] and Jianwu Dang[3, e]

[1]Tianjin key Laboratory of Cognitive Computation & its Applications, School of Computer Science, Tianjin University, China

[2]Department of Computer Science and Technology, Tsinghua University

[3]School of Information Science, Japan Advanced Institute of Science and Technology, Japan

[a]chensai@tju.edu.cn, [b]hcwang@tju.edu.cn, [c]jjia@tsinghua.edu.cn, [d]ayt@tju.edu.cn, [e]jdang@jaist.ac.jp

*The corresponding author

Keywords: PLP; MFCC; perceptual measurement; acoustics; Chinese initials.

Abstract. Many works have been done in the methods of improving performance by proposing new speech characteristics and new perception measurements. However, they only focus on one of the two aspects. In this paper, we try to study the relationship between them. That is, we discuss which acoustic features or their combinations are the most consistent with the real perception of Chinese initials. We propose a method that can measure the acoustic distance and keep it monotonically related to the perceptual distance of Chinese initials. We first define the acoustic distance and perceptual distance between different Chinese initials, and single out a proper combination of acoustic features and two compatible distance metrics by conducting clustering analysis on the samples of all types of Chinese initials using MFCC and PLP. Based on the data provided by the General Hospital of the People's Liberation Army, we then calculate the acoustic distance and perceptual distance. Finally, we calculate the Spearman's rho between two types of distance corresponding to the two calculation method. The experiment results show that there is a relatively high strength of monotonic relationship with the selected acoustic features between two types of distance.

Introduction

In traditional phonology, the place and the manner of the articulation in the vocal tract are applied to classify Chinese initials. And the statistical and psychological methods are used to explore the perceptual characteristics. 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 [1, 2]. A perceptual measurement based on LPC among Chinese finals has been proposed 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.

In this paper, we discuss which acoustic features or their combinations are the most consistent with the perception of Chinese initials. We systematically test PLP and MFCC representations of Chinese initials by carrying out two experiments with respect to acoustic space and perceptual space, respectively. We then combine the results of the two experiments by using a statistical method, called Spearman's rank correlation coefficient, to assess how well the relationship between two types of

distance can be described using a monotonic function. We also single out a proper acoustic feature representations for Chinese initials and distance metrics between different categories of initials to measure the acoustic distance which is monotonically related to the perceptual distance.

Acoustic Experiment and Results

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 [6], 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 categories of initials as the distance between two clusters. We use hierarchical methods because it leads to smaller computation costs by not having to worry about a combinatorial number of different choices [7], which is suitable for the task attempting to use as many dissimilarity measures as possible.

Mel Frequency Cepstrum Coefficient versus Perceptual Linear Predictive. There are many similarities between MFCC and PLP [8, p. 67]. Fig. 1 shows a comparative scheme of PLP and MFCC computation. Differences between PLP and MFCC lie in the filter-banks, the equal-loudness pre-emphasis, the intensity-to-loudness conversion and the application of LP, each of which makes PLP more consistent with human auditory impression [9].

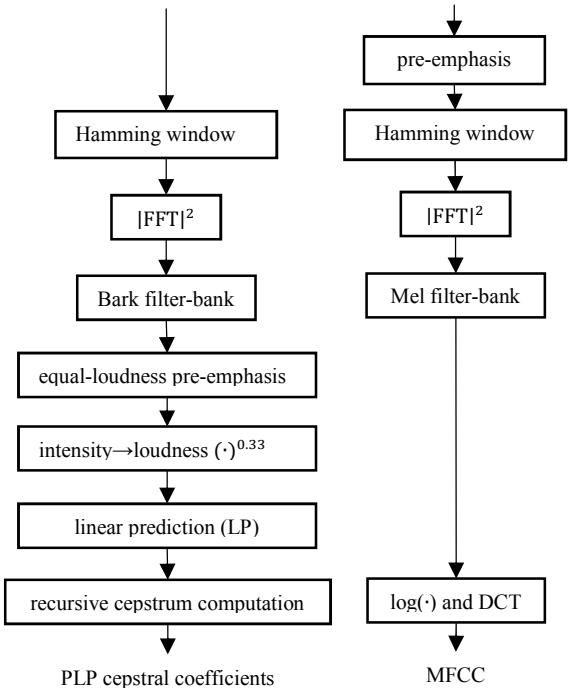

Figure 1. The computation steps of PLP (left) and MFCC (right).

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 is 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) [10], 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 [11] are used in this paper. They are listed in table 2.

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 labelled 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.

Clustering Analysis and Results. We extract 12 coefficients of MFCC for each frame of an initial, and calculate the mean of coefficients of all the frames as the 12 coefficients of the initial. The 12 coefficients of PLP of an initial are obtained using the same method. Both MFCC and PLP coefficients were calculated using the rastamat Matlab toolbox [12] with parameters that resemble feature extraction from the HTK software [13], i.e. the frame length is 25 milliseconds and the moving step is 10 milliseconds. The highest band edge of filters is 8000Hz and the number of warper spectral bands to use is 22.

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 one of the 12 coefficients A, respectively.

We generate all possible combinations of 12 MFCC and PLP coefficients respectively (the total number of possible combinations of MFCC and PLP coefficients tots up to 4095 respectively). We define the accuracy of hierarchical clustering of the initials, Acc, as follows:

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

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 12 MFCC or PLP coefficients and all dissimilarity metrics (43 in total), and then calculate the average accuracy of hierarchical clustering of

initials, $\overline{\text{Acc}}$, which is defined as the arithmetic mean of 21 Acc corresponding to the 21 categories of initial. We expect $\overline{\text{Acc}}$ 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 Chebyshev has the highest $\overline{\text{Acc}}$. We can also see the clustering using PLP has higher $\overline{\text{Acc}}$ than those using MFCC for all the exponents (i.e. ∞, 10, 5, 4, 3, 2, 1). Hence, we can infer that PLP is more consistent with the perception of Chinese initials than MFCC, and the Shortest and Chebyshev are the most compatible distance metric as the inter-cluster and intra-cluster metrics, respectively, with both PLP and MFCC.

TABLE III. AVERAGE ACCURACY OF HIERARCHICAL CLUSTERING USING MFCC AND PLP

Distance between Clusters–Dissimilarity of Objects	**PLP**		**MFCC**	
	$\overline{\text{ACC}}$	*Var*	$\overline{\text{ACC}}$	*Var*
Shortest–Chebyshev(exp=∞)	0.9365	0.0010	0.9304	0.0008
Shortest–Minkowski(exp=10)	0.9352	0.0011	0.9293	0.0008
Shortest–Minkowski(exp=5)	0.9344	0.0011	0.9277	0.0008
Shortest–Minkowski(exp=4)	0.9342	0.0011	0.9271	0.0008
Shortest–Minkowski(exp=3)	0.9335	0.0011	0.9270	0.0008
Shortest–Euclidean(exp=2)	0.9323	0.0011	0.9269	0.0009
Shortest–Std.Euclidean(exp=2)	0.9321	0.0011	0.9264	0.0009
Shortest–Manhattan(exp=1)	0.9312	0.0011	0.9277	0.0008
Shortest–Cosine	0.9077	0.0119	0.9082	0.0153
Shortest–Correlation	0.8633	0.0307	0.8659	0.0366

Perceptual Experiment and Results

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

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

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 [14]. 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 $\mathrm{e(i,j)}$ 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 $\mathrm{p(i,j)}$ indicates the perceptual distance between the ith initial and the jth initial.

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:

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

We have inferred that PLP is more consistent with the perception of Chinese initials than MFCC, and the Shortest and Chebyshev are the most compatible distance metrics between clusters and samples respectively, with both PLP and MFCC. In order to verify this assumption, we calculate the Spearman's rho using each of the 4095 acoustic distance matrices and the perceptual distance matrix. We then single out the maximum of the 4095 Spearman's rho and the dimensions of 12 PLP coefficients corresponding to that maximum. The maximum Spearman's rho using MFCC and the same distance metric is also calculated in comparison with that using PLP.

Because the UPGMC and Euclidean, as an inter-cluster and intra-cluster distance respectively, are considered as being perceived directly through the senses in perceptual measurement of Chinese initials [2], we also calculate the maximum Spearman's rho using UPGMC and Euclidean in comparison with the Shortest and Chebyshev for both MFCC and PLP. The results are shown in table 4.

Table 4 shows that a Spearman correlation of 0.6328 occurs when the acoustic distance is computed using PLP (corresponding to the dimensions 4th, 5th, 8th, 9th, 10th, 11th) with the Shortest and Chebyshev as the inter-cluster and intra-cluster distance metrics, respectively. It is larger than any other result in the table, which essentially agrees with what we have inferred in Section E that PLP is more consistent with the perception of Chinese initials than MFCC. The results also show that the Shortest and Chebyshev are the most compatible inter-cluster and intra-cluster distance metrics, respectively, with both PLP and MFCC. Note that even though the Spearman's rho, 0.6328, is not extremely large, we can also conclude that there exists a high strength of monotonic relationship because we invite only 20 subjects (the number of subjects is not extremely large) to participate in the experiment, which doesn't make the confusion matrix exactly precise due to the small probability events caused by subjects' absent-mind, weariness, or the equipment's inaccuracy. We eliminate those errors by ignoring the elements less than or equal to 0.01 in confusion matrix, which we have discussed in Part III. However, when we only have the largest 10 elements of the confusion matrix in reserve, we obtain the maximum Spearman's rho that is equal to 1 using PLP with the Shortest and Chebyshev, which means the acoustic distance is definitely a perfect monotone function of the perceptual distance.

TABLE IV. SPEAMAN'S RHO USING MFCC AND PLP

Feature Representation	Distance between Clusters	Dissimilarity of Objects	Spearman's rho
MFCC	Shortest	Chebyshev	0.5727
MFCC	UPGMC	Euclidean	0.5644
PLP	Shortest	Chebyshev	0.6328
PLP	UPGMC	Euclidean	0.5835

Conclusions

In this paper, we discuss which acoustic features or their combinations are the most consistent with the perception of Chinese initials. The PLP and MFCC representations of Chinese initials are systematically tested by carrying out two experiments with respect to acoustic space and perceptual space. The experimental results show that the acoustic distance using PLP is more monotonically related to the perceptual distance of Chinese initials than that using MFCC, which means that PLP is more consistent with the perception and more suitable for perceptual measurement of Chinese initials than MFCC. The results also show that the Shortest and Chebyshev are the most compatible distance metrics between clusters and samples, respectively, with both PLP and MFCC. Besides, we single out a proper combination of acoustic features and two compatible distance metrics, which can be used to automatically evaluate whether the audiometric word lists are equivalent to each other.

Acknowledgements

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 and No.6117501.

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] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2006.

[7] Johnson, Stephen C. "Hierarchical clustering schemes." Psychometrika 32.3 (1967): 241-254.

[8] E. Schukat-Talamazzini, Automatische Spracherkennung–Grundlagen, statistische Modelle und effiziente Algorithmen.Braunschweig: Vieweg, 1995.

[9] Hönig, Florian, et al. "Revising perceptual linear prediction (PLP)." Proceedings of INTERSPEECH. 2005.

[10] Deza, Michel Marie, and Elena Deza. Encyclopedia of distances. Springer Berlin Heidelberg, 2009.

[11] Murtagh, Fionn. "Complexities of hierarchic clustering algorithms: State of the art." Computational Statistics Quarterly 1.2 (1984): 101-113.

[12] 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/

[13] 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.

[14] Boothroyd, Arthur. "The performance/intensity function: an underused resource." Ear and hearing 29.4 (2008): 479-491.