A SYSTEMATIC APPROACH TO COMPUTE PERCEPTUAL DISTRIBUTION OF MONOSYLLABLES

Yuhao Wu, Jia Jia, Feng Lu, and Lianhong Cai

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology(TNList)
Email: {yh-wu14@mails, jjia@mail, lu-f13@mails, clh-dcs@mail}.tsinghua.edu.cn

ABSTRACT

In speech understanding, perceptual computing is widely used to quantify the perceptual properties. Previous researches on perceptual computing mainly focused on the level of phonemes (i.e. consonants and vowels). However, perceptual measurement in the level of syllables is also needed in scenarios such as speech recognition. To tackle this problem, we propose a systematic approach to calculate the perceptual distribution of monosyllables. It is composed of three parts. First, we generate a feature vector from each monosyllable based on acoustic property. Second, we construct the perceptual space based on the perception distance of every two feature vectors. Third, we measure the perceptual distribution for these monosyllables based on the perceptual space and a constraint matrix. Experiments show that 1) cluster results are in accordance with articulation position category in acoustics, 2) recognition rate of audiometry is within the standard range of performance-intensity function, 3) distribution of paracusia is consistent with the computation results of perceptual distribution.

Index Terms— Speech perception, perceptual distribution, acoustic features, monosyllable signals

1. INTRODUCTION

Perceptual properties of speech signals are the critical part of speech understanding [1, 2, 3, 4]. Conventionally, the perceptual properties are concluded from a great deal of clinical study such as the recognition test [5, 6, 7]. Recently, perceptual computing of speech is introduced to provide a quantified and convenient approach to measure the perception of speech [8, 9]. It extracts the acoustic features and then computes perceptual distribution.

Previous researches on the perceptual computing of speech mainly focused on the level of phonemes (i.e. consonants and vowels). They extracted time domain and frequency domain acoustic features to compute articulation index of consonants [10, 11]. Moreover, they obtained the perceptual categories of vowels from processing the frequency domain features and enhanced the tone perception based on fundamental frequency envelope [12, 13, 14].

However, perceptual measurements in the level of syllables is sometimes needed as well. For example, clinical speech recognition tests contain the test on syllable recognition [15]. Moreover, the monosyllable is the basic unit of daily speech rather than the phoneme and thereby perceptual computing in the level of monosyllables potentially enhances the validity of the speech perception measurement [16]. Additionally, perceptual distribution of syllables affects the discrimination of corpus, and further affects the validity and reliability of clinical audiometry [17]. Therefore, it is vitally important to research the perceptual distribution of monosyllables.

To measure the perceptual distribution of monosyllables, we propose an systematic approach. It uses the monosyllabes as input and generates perceptual distribution of these monosyllables. The basic intuition of our approach is that an intermediate representation is needed to map the acoustic physical property to perceptual property, which we call the perceptual space of monosyllables. The original syllables are processed into the perceptual space and then the perceptual space can be used to derive the perceptual distribution. Our approach is composed of three parts. The first part is responsible for feature vector generation of each monosyllable, which is based on the acoustic property and the second part is for the perception space construction, which is based on the perception distance of every two feature vectors. The last part measures the perceptual distribution with the perception space and a constraint matrix. Contributions of this paper can be concluded as follows:

- We propose a systematic approach to calculate the perceptual distribution of monosyllables. Experimental results show that recognition rate of audiometry is within the standard range of performance-intensity function, and that distribution of paracusia is consistent with the computation results of perceptual distribution.
- We introduce the feature vectors of a monosyllable, which represent its acoustic features. Cluster results

are in accordance with articulation position category.

- We design a method to construct perceptual space of monosyllables based on perceptual distance. It can be further applied to all the syllabic corpus.

2. AN OVERVIEW OF THE SYSTEMATIC APPROACH

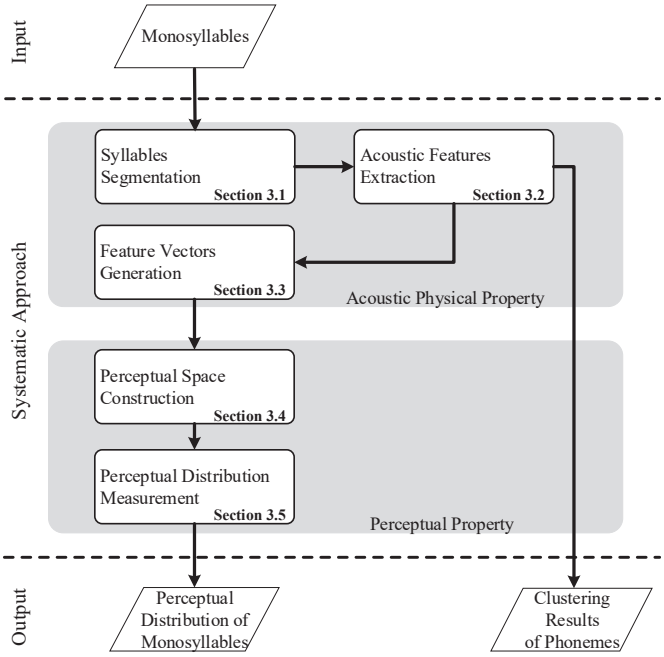

Fig. 1. An overview of the systematic approach to compute the perceptual distribution of monosyllables.

Fig.1 illustrates an overview of the our approach to measure perceptual distribution of monosyllables. During the pre-processing, monosyllables are first segmented into consonants and vowels (Section 3.1) and then the acoustic features including the time-domain and frequency-domain are extracted (Section 3.2). With the weighted summary of the features of the phonemes, feature vectors of the monosyllable are generated, which represent its acoustic property (Section 3.3). Provided with the feature vectors, the perceptual distance of each two monosyllables are calculated and then the perceptual space is constructed, which represents the confusion degree of these monosyllables (Section 3.4). The perceptual distribution is measured based on the perceptual space (Section 3.5). Clusters of phonemes (consonants and vowels) are intermediate results to validate the accuracy of acoustic features and perceptual distance. The perceptual distribution of the monosyllables is the output of our approach.

3. THE DESIGN AND IMPLEMENTATION OF THE SYSTEMATIC APPROACH

We choose Chinese mandarin monosyllabic corpus as a representative target in this work. Our purposed approach constructs its perceptual space, and measure the perceptual distribution of the monosyllables in the corpus.

3.1. Syllables Segmentation

A monosyllable contains a consonant, a vowel, and a tone. Consonants and vowels have different perceptual features, so we segment the monosyllables into consonants and vowels separately [18]. During the segmentation, signal length, amplitude and energy are recorded for each monosyllable.

3.2. Acoustic Features Analysis of Consonants, Vowels and Tones

Based on the time-domain and frequency-domain acoustic characteristic of speech signals, we extract the vector features of consonants, vowels and tone, which contains 38, 9 and 6 dimensions respectively.

The unit and magnitude of the elements in feature vectors are usually greatly different, so we normalize features with Equ 1, which guarantees the value of features are within the range of [0, 1]. It eases the construction of perception space.

$$x_{new} = \frac{x_{pre} - \min(X_{pre})}{\max(X_{pre}) - \min(X_{pre})} \quad (1)$$

In Equ 1, x_{new} represents the normalized feature vectors. x_{pre} denotes the extracted feature vectors, and X_{pre} denotes the set of x_{pre}. It is worth to be noted that each element of the feature vector is normalized independently.

3.2.1. Acoustic Features of Consonants

Generally, signal length and energy of a consonant is short and weak. Human ears are sensitive to acoustic characteristics in frequency domain. So we extract a 38-dimension feature vector, which is composed of 5 dimensions for time domain [19], 12 dimensions for Mel Frequency Cepstral Coefficients (MFCC), and 21 dimensions for bark band coefficients [20].

3.2.2. Acoustic Features of Vowels

A vowel follows a consonant in general situations. Therefore, vowels perception should be considered with transition condition from consonants [13]. So we choose linear predictive coding (LPC) as acoustic features of vowels.

Taking 500 Hz, 1000 Hz, 2000 Hz as center, computing the intergral of the LPC coefficients at bands of [450, 550] Hz, [950, 1050] Hz, and [1950, 2050] Hz respectively, we get a group of 9-dimension vectors as the features of vowels.

	PD Average of intra-Class	PD Average of inter-Classes	Case of intra-Class	Case of inter-Classes
Consonants	0.396	0.796	0.232 (j/q/x)	1.174 (j/q/x ↔ l/m/n/r)
Vowels	0.387	0.820	0.269 (ia/iang/iao)	1.023 (ia/iang/iao ↔ ui/un)

Table 1. Perceptual Distance of Cluster Results. The perception distance values of inter-classes are much higher than those of intra-class.

3.2.3. Acoustic Features of Tones

Tone plays an important role in perception of speech signals. The main difference between tones lies in fundamental frequency [14]. So we set fundamental frequency as the acoustic features of tones.

We compute values of fundamental frequency with Equ 2, where $f0_n$ is fundamental frequency of the n-th sampling point, $f0_{pre}(n)$ is fundamental frequency to be processed, $F0_{mean}$ is the mean value of all the fundamental frequencies before being processed, and F is constant of the slope of the f0 contour.

$$f0_n = F \times (f0_{pre}(n) - F0_{mean}) + F0_{mean} \quad (2)$$

3.3. Feature Vectors of Monosyllables

Consonants, vowels and tones all affect each other for the speech perception of the monosyllables. For syllables with the same vowel but different consonants, acoustic perception will change at the starting position of vowels. For example, for the same [ang] in [bang] and [cang], LPC coefficients of the first three dimensions are different because of the difference in consonants [b] and [c]. And for syllables with the same consonant but different vowels, perception condition changes at the ending position of consonants. For example, for the same [b] in [bang] and [bing], f0 contours are different because of the difference at the starting position of [ang] and [ing]. Tones have a great influence on the ending position of vowels. For the same [bang] with the first tone (bāng) and the falling-rising tone (bǎng), LPC coefficients of the last three dimensions are different.

Based on the above reasons, independently measuring phonemes cannot fully represent perceptual condition of speech signals. So we propose feature vectors in the level of monosyllable as described in Equ 3.

$$S = \alpha C + \beta V + \gamma T \quad (3)$$

In Equ 3, S represents the feature vector of a monosyllable, while the C, V, T represents the feature vector of the consonant, the vowel, and tone respectively.

In feature vectors of monosyllables, acoustic features of consonants (α), vowels (β), and tones (γ) are parameterized with different weights. In a monosyllable, the signal length and energy of a consonant is much shorter and lower than a vowel. So the weights of consonants should be lower than those of vowels. Since the tone has an impressive impact on the pronunciation of the vowel but not on the consonant, so the weight of tones should be within the range of consonants and vowels [8]. We vary the three parameters (α, β, γ) for many monosyllabic corpus to find the best fitting values. They are 0.10, 0.65, 0.25 respectively.

3.4. Perceptual Space of Monosyllables based on Perceptual Distance

Before computing perceptual space of monosyllables, we need to determine the method to measure perception distance between two syllables. The euclidean distance is intuitive, which reflects the human perception condition [8]. In our work, we measure perception distance between syllables using euclidean distance of their feature vectors.

We apply the hierarchical clustering independently on the consonants and vowels based on perceptual distance. Comparing the cluster results with the speech articulation position, the accuracy of acoustic feature vectors can be figured out [1].

From the cluster results, we can see that there are 9 classes of consonants (i.e. {j/q/x}, {ch/sh/zh}, {c/s/z}, {h/k/p/t}, {l/m/n/r}, {b/d/f/g}, {y}, {w}, {zero consonant}) and 9 classes of vowels (i.e. {o/uo/e/eng/ou/ong/ueing/u}, {ui/un}, {ia/iao/iang}, {iong/iou/iu}, {a/er/ang/ao}, {ua/uang/uai/uan}, {ai/an}, {ei/ü/ün/üan/üe}, {en/i/ian/ing/ie/in}). Compared to the speech articulation position, the accuracy of clustering results is validated [1].

We compute the perceptual distance of every two monosyllables in each class, and then set the average value as its center point. Perceptual distance of the two center points is defined as the perceptual distance between the two classes. The perception distance values of inter-classes are much higher than those of intra-class, as shown in Table 1.

3.5. Perceptual Distribution of Monosyllables

For each monosyllable, we select three monosyllables closest to it from the perceptual space and the selection algorithm is shown in Equ 4. This is because there are three phonemes(i.e. consonants, vowels, tones) in a monosyllable. These four

	j/q/x	ch/sh/zh	c/s/z	h/k/p/t	l/m/n/r	b/d/f/g	y	w	zero
o/uo/e/eng/ou/ ong/ueng/u	0.488	0.521	0.458	0.461	0.443	0.449	0.464	0.467	0.470
ui/un	0.494	0.527	0.464	0.467	0.449	0.455	0.470	0.473	0.476
ia/iang/iao	0.533	0.566	0.503	0.506	0.488	0.494	0.509	0.512	0.515
iong/iou/iu	0.512	0.545	0.482	0.485	0.467	0.473	0.488	0.491	0.494
a/er/ang/ao	0.530	0.563	0.500	0.503	0.485	0.491	0.506	0.509	0.512
ua/uang/uai/uan	0.506	0.539	0.476	0.479	0.461	0.467	0.482	0.485	0.488
ai/an	0.452	0.485	0.422	0.425	0.407	0.413	0.428	0.431	0.434
ei/ü/ün/üan/üe	0.452	0.485	0.422	0.425	0.407	0.413	0.428	0.431	0.434
en/i/ian/ing/ie/in	0.362	0.395	0.332	0.335	0.317	0.323	0.338	0.341	0.344

Table 2. The Constraint Matrix of Perceptual Space.

monosyllables are defined as one perceptual group, which is the minimum unit of the perceptual space.

$$\begin{cases} \bar{p_i} = P \pm \delta \\ |\bar{p_i} - p_i| \leq \epsilon \end{cases} \tag{4}$$

In Equ 4, p_i is the perception distance between test syllable and candidate syllables, $\bar{p_i}$ is mean value of all the items to be selected, and P is a constraint constant for the perceptual group. Considering perceptual distance of phonemes from one or two classes is not always the same, we set a unique constraint value (P) for each perceptual group. The value of P depends on the main monosyllable of each perceptual group. Through a survey on two corpus (males and females voice), we find the best fitting values of δ, ϵ and P. The δ and ϵ are both 0.1, and P is a constraint matrix shown in Table 2.

4. EXPERIMENTAL RESULTS

We take a male-voice mandarin monosyllabic corpus as our experimental target, which contains 1251 monosyllables and covers almost all of our daily characters. The monosyllables are segmented and tagged with VisualSpeech [18].

We conduct an audiometry experiment, using the perceptual groups in Section 3.5 as test materials. We set one monosyllable as test item, together with three monosyllables in the same perceptual group as confusion items. 30 volunteers (22 males and 8 females) from different background participated the speech signals recognition tests, which were generated according to the test groups. Each participant took 60 perceptual groups in the experiment. The average recognition rate is 82.1% and the variance is 6.7%, which is within the standard speech recognition range[21, 22]. It indicates the accuracy of the generated feature vectors of the monosyllables.

From the results of wrong recognition, 69.4% of the confusion recognition come from consonants, while 26.5% come from vowels confusion and the other 4.1% are from tones. Further more, in the consonants confusion, 92.9% confusion items are in the same perceptual group with the test syllable. For the vowels confusion, 89.1% of the confusion items are in the same perceptual group with the test syllable. We can see that distribution of paracusia is consistent with the computation results of perceptual distribution.

From the experimental results, we can conclude that, 1) recognition rate of audiometry is within the standard range of performance-intensity function, validating the accuracy of generated acoustic features in Section 3.3; 2) distribution of paracusia is consistent with the computation results of perceptual distribution, indicating the validity and reliability of measured perceputal distribution. More comparison experiments will be conduct in our future work.

5. CONCLUSIONS

In this paper, we present a systematic approach to measure perceptual distribution of monosyllables. We introduce the feature vectors, which represent the acoustic features of a monosyllable. In addition, based on the perceptual distance, we purpose the perceptual space of monosyllables to map the acoustic property to perceptual property. Clustering results of phonemes are in accordance with articulation position category in acoustics, which validates the accuracy of the extracted acoustic features. The recognition rate of audiometry is within the standard range of performance-intensity function and distribution of paracusia is consistent with the computation results of perceptual distribution, which indicate the validity and reliability of measured perceputal distribution. The idea of perception distance can be applied to other languages.

6. ACKNOWLEDGEMENTS

This work is supported by National Key Research and Development Plan (2016YFB1001200), the innovation method fund of China (2016IM010200), the National Basic Research Program (973 Program) of China (2013CB329304), National Natural, and Science Foundation of China (61370023, 61602033). This work is partially supported by the major project of the National Social Science Foundation of China (13ZD189).

7. REFERENCES

[1] Hisao Kuwabara, "Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate.," in *EUROSPEECH*, 1997.

[2] Michael Blomgren, Yang Chen, Manwa L Ng, and Harvey R Gilbert, "Acoustic, aerodynamic, physiologic, and perceptual properties of modal and vocal fry registers," *The Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2649–2658, 1998.

[3] Steven Greenberg, "Understanding speech understanding: Towards a unified theory of speech perception," in *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*. Keele, England, 1996, pp. 1–8.

[4] Sarah Hawkins, "Roles and representations of systematic fine phonetic detail in speech understanding," *Journal of Phonetics*, vol. 31, no. 3, pp. 373–405, 2003.

[5] Johan't Hart, René Collier, and Antonie Cohen, *A perceptual study of intonation: an experimental-phonetic approach to speech melody*, Cambridge University Press, 2006.

[6] Catherine T Best, Gerald W McRoberts, and Nomathemba M Sithole, "Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by english-speaking adults and infants.," *Journal of experimental psychology: human perception and performance*, vol. 14, no. 3, pp. 345, 1988.

[7] David B Pisoni and Christopher S Martin, "Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses," *Alcoholism: Clinical and Experimental Research*, vol. 13, no. 4, pp. 577–587, 1989.

[8] DeLiang Wang and Guy J Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*, Wiley-IEEE Press, 2006.

[9] Dennis Norris, James M McQueen, and Anne Cutler, "Perceptual learning in speech," *Cognitive psychology*, vol. 47, no. 2, pp. 204–238, 2003.

[10] Zhang Jialu, "Distinguishing feature system of mandarin chinese," *ACTA ACOUSTICA*, vol. 30, no. 6, pp. 506–514, 2005.

[11] Zhang Jialu, "Distinctive system tree of chinese mandarin," *ACTA ACOUSTICA*, vol. 31, no. 3, pp. 193–198, 2006.

[12] James Emil Flege, Ocke-Schwen Bohn, and Sunyoung Jang, "Effects of experience on non-native speakers' production and perception of english vowels," *Journal of phonetics*, vol. 25, no. 4, pp. 437–470, 1997.

[13] Peter F Assmann and Quentin Summerfield, "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," *The Journal of the Acoustical Society of America*, vol. 88, no. 2, pp. 680–697, 1990.

[14] Denise Klein, Robert J Zatorre, Brenda Milner, and Viviane Zhao, "A cross-linguistic pet study of tone perception in mandarin chinese and english speakers," *Neuroimage*, vol. 13, no. 4, pp. 646–653, 2001.

[15] Aravind Ganapathiraju, Jonathan Hamaker, Joseph Picone, Mark Ordowski, and George R Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on speech and audio processing*, vol. 9, no. 4, pp. 358–366, 2001.

[16] James M Pickett, *The acoustics of speech communication: Fundamentals, speech perception theory, and technology*, Allyn & Bacon, 1999.

[17] Martin Cooke and Daniel PW Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech communication*, vol. 35, no. 3, pp. 141–177, 2001.

[18] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 31–35.

[19] Metin Akay, *Time Frequency andWavelets in Biomedical Signal Processing*, IEEE Press series in biomedical Engineering, 1998.

[20] Antti Eronen and Anssi Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*. IEEE, 2000, vol. 2, pp. II753–II756.

[21] Martin Cooke, Phil D Green, and Malcolm Crawford, "Handling missing data in speech recognition.," in *ICSLP*, 1994.

[22] Annamaria Mesaros, Toni Heittola, Onur Dikmen, and Tuomas Virtanen, "Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 151–155.