



WESPAC IX 2006

The 9th Western Pacific Acoustics Conference
Seoul, Korea, June 26-28, 2006

A NEW TIME-DOMAIN FEATURE PARAMETER FOR PHONEME CLASSIFICATION

Shuiyuan YU, Donghong ZHENG, Xiaoliang FENG, Yudong CHEN

Communication Acoustics Laboratory

Communication University of China

Beijing 100024, P.R.China

E-mail: yusy@cuc.edu.cn

hong-02@163.com

Banban220@gmail.com

bychenyudong@cuc.edu.cn

ABSTRACT

In this paper a new time-domain feature parameter is presented for phoneme classification. The parameter results from reconstructed phase space of original speech signal. Isolated phoneme classification using features from time-domain phase space reconstruction has been investigated recently, however competent representation of phoneme class feature vectors has not been presented up to now. Our parameter expresses the component structural characters of speech signal. An experiment is designed using the parameter presented by us for phoneme classification and the results show that the parameter is effect to classify phoneme.

KEYWORDS: Phase space, classification phoneme

I. INTRODUCTION

Phase space reconstruction techniques are considered as an effective method to capture nonlinear and higher-order characteristics of the speech. Reconstructed phase spaces (RSP) are topologically equivalent to the original system, if the embedding dimension is large enough (Sauer et al., 1991).

Given a time series $X=\{x_t, t=1,2,\dots,N\}$, where t is a time index, and N is the number of observations, a reconstructed phase space is formed, according to [1],

$$\vec{X} = (x_{t-(m-1)\tau}, \dots, x_{t-2\tau}, x_{t-\tau}, x_t) \quad (1)$$

Where τ is the time delay and m is the embedding dimension. Although time delay and embedding dimension are important reconstructed phase space parameters, they have not been extensively studied in this preliminary work on phoneme classification, and many researchers used empirical values. In this paper, we use an embedding dimension of 2 and a time delay of 6. about the detailed study on the lag and dimension parameters of the reconstruction process for speech can be found in[2].

Examples of \vec{X} trajectory for 2 different phonetic classes are shown in Fig. 1.

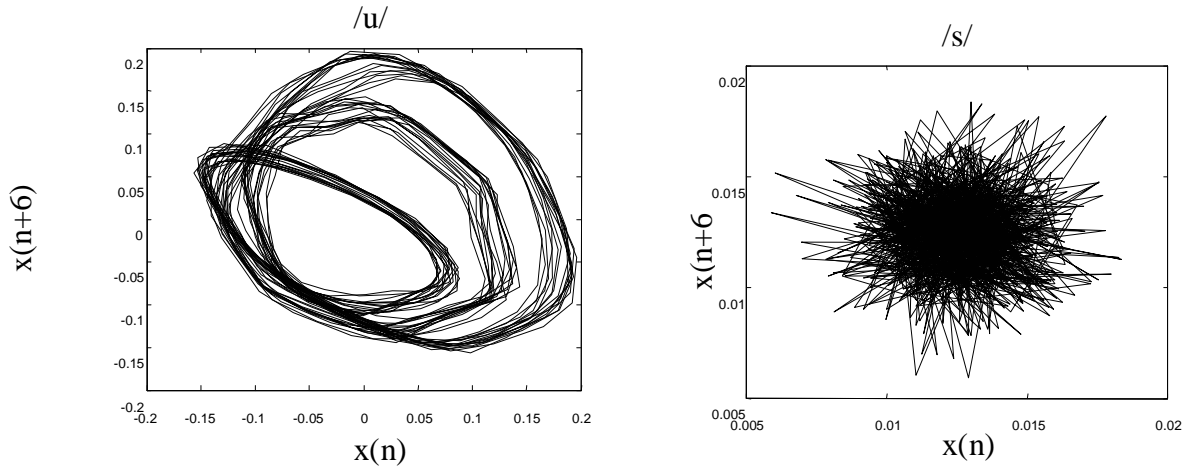


Fig.1 RSP trajectory of ‘/u/’ phoneme (left), and ‘/s/’ phoneme (right).

Fig.1 gives a portrait of the reconstructed phase space trajectory. In reconstructed phase space, some structural features which are illegible in time domain waveform become obvious. Vowel has a fundamental frequency, which exhibits a number of homocentric circles, while the trajectory of fricative in RPS is very different.

In order to give expression to the pattern of phonemic trajectory in RPS, many researchers proposed many different approaches. From nonlinear dynamics, [3][4] used fractal dimensions and Lyapunov exponents. [5] directly quantified the graph of phonemic trajectory in RPS, they divided the reconstructed phase space into 100 histogram bins, and then computed natural distribution of the attractor, finally applied Principal Component Analysis (PCA) to the distribution. [6] used 4 features to express the chaotic features: the mean of the correlation sum and its the standard deviation, and the mean of the scale-varying correlation dimension and its standard deviation. In order to improve the performance, they also incorporated the four features into mel-frequency cepstrum coefficients (MFCC). [7] and [1] used a GMM to capture the characteristic attractor structure of speech phonemes. [8] used global flow reconstruction to

generate a compact and quantitative description of the structure and trajectory of vowel attractors in a reconstructed phase space. In order to improve robustness of automatic speech recognition (ASR) systems to noise, in [9] a sub-band approach was used, RPS was set in sub-band filtered signal. These methods attempted to capture patterns of phonemic trajectory in RPS from many aspects.

In this paper, we explore a novel time-domain approach to modeling and classifying speech phoneme waveforms. The goal of our work is to directly model reconstructed phase spaces for phoneme classification. Our approach is very simply computationally.

This paper is organised as follows. In section II, we propose a new time domain vector, called sonant exponent (SE). Section III extracts SE values of 21 phonemes from 100 females, and plots them in one space, which is convenient to discuss SE's performance of representing phonetic features of phonemes and the relation between phonemes (including vowels and consonants). Section IV introduces an experiment to test the ability of this new time domain feature proposed by us for phoneme classification. Section V summarizes this new approach.

II. A NEW TIME-DOMAIN FEATURE PARAMETER

Each class of phoneme has its own acoustic character. A reconstructed phase space (RPS) is an embedding that maps a signal into a sufficiently high dimension. The RPS is constructed typically by mapping time lagged copies of original signal onto axes of the new high dimensional space. In the new high dimensional space a structure is formed that is topologically equivalent to the original phase space. A speech signal has its own specific structure in the new high dimensional space. A periodic signal shows a circular structure, and a noise forms a jumbled mass, so the vowel presents a quasi-circular structure and the fricative exhibits a disorderly straw. Our basic idea is to design variables to represent the degree of circle and chaos.

First, we define a variable, called scalar displacement (abbreviated as D_s), to express total trajectory length of a speech signal in its RPS. It is given by:

$$D_s = \sum_{n=1}^N \left| \vec{x}_{n+1} - \vec{x}_n \right| \quad (2)$$

Due to its non-continuity, we expect that a fricative has a more D_s value than a vowel.

Second, we define another variable, called radial displacement (abbreviated as D_r), to express accumulative total of radial displacements of a speech signal's trajectory in its RPS, given by:

$$D_r = \sum_{n=1}^N \left| \left| \vec{x}_{n+1} \right| - \left| \vec{x}_n \right| \right| \quad (3)$$

That is, a circle has a D_r value of zero in RPS. Due to its quasi-circular trajectory in RPS, we

expect that a vowel has a small D_r value than a fricative.

D_s and D_r represent structural feature of a speech in its RPS from different aspects. D_s represents the degree of chaos, and D_r represents the degree of periodicity. Now we define a 2-dimension variable, called sonant exponent (abbreviated as SE in rest sections). We expect that SE can reflect phonetic feature of a speech signal. SE is given by incorporating D_s and D_r .

$$SE = [D_s, D_r] \quad (4)$$

III. SE SPACE AND ITS PHONETIC FEATURE

We extracted SE values of 21 Mandarin phonemes from 100 female speeches. They are 6 vowels: /a/, /i/, /u/, /o/, /y/ and /ɤ/, 15 consonants: /ts'/, /s/, /ts/, /tɕ'/, /ɕ/, /tɕ/ /tɕ/, /ç/, /tɕ'/, /r/, /x/, /f/, /l/, /m/ and /n/. These SE feature vectors are typically computed over a 10 ms window. Fig.2 shows positions of these 21 phonemes in SE space. It should be noticed that positions of several phonemes almost overlap each other, they are /y/, /a/, /r/ and /m/, /n/.

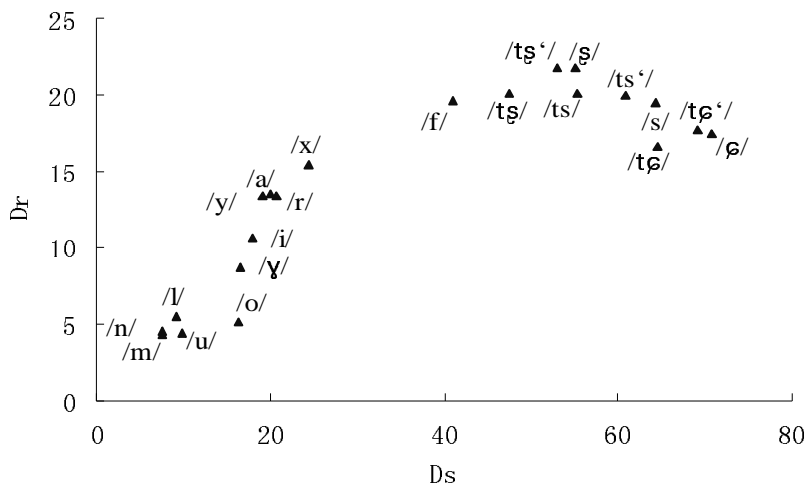


Fig.2 The distribution of 21 Mandarin phonemes in SE space

In SE plan, we take notice of that these positions show themselves into 3 groups: /m/ /n/ /l/ /u/ are first group, vowels /ɤ //i/ /y/ /a/ /r/ are second group, and /tɕ/ / tɕ'/ / ɕ/ /ts/ /ts'/ /s/ /tɕ/ /tɕ'/ /ç/ are third group. /o/, /x/ and /f/ are lone respectively. In group 1, besides vowel /u/ the rest are nasal and lateral, in group 2, besides approximant /r/ the rest are vowels, and the group 3 includes all of affricate and fricative, and in this group laminal, apical and retroflex compose subgroups respectively. Of these 21 phonemes, several phonemes are peculiar: /x/ is radical fricative, but near vowel group. Vowel /u/ is near nasal and lateral group ,and retroflexed approximant /r/ is mixed with vowel group. That is to say, phonemes which have similar

phonetic quality, taking one with another, are closer in SE space.

In Fig.2, we notice that in all of 3 affricate and fricative groups (retroflex, apical and laminal group), D_s values of fricative are the largest, D_s values of aspiration are middle, and D_s of unaspiration are the least. In all 21 phonemes, D_r values of affricate and fricative are the largest, ones of vowel are middle, and ones of nasal and lateral are the least. There are small variational range of D_s and D_r values in nasal and lateral group, and small variational range of D_s value and large variational range of D_r value in vowel group, and there are large variational range of D_s and small variational range of D_r in affricate and fricative group.

IV. EXPERIMENT

In order to test the effect on classification phoneme using SE we use 15 consonants from other 96 females, these phonemes are embedded into monosyllable respectively when they are pronounced, and segmented artificially. The Phoneme Classification algorithm used in our experiment is following: first, speeches are divided into frames of 10ms duration; second, SE value is computed for each frame in light of equation (2), (3), and (4); third, SE values are averaged over all frames of each speech, and the mean is treated as SE value of this speech; finally, we compare the SE value of each speech with each templete value in Fig. 2, the phoneme label with the small distance is regarded as the classification result. Table 1 shows the all classification results.

Table 1 -accuracy of classification phoneme

phoneme	accuracy	Accuracy over group
/ts'/	16%	39%
/tʂ'/	26%	62%
/tʃ/	14.6%	44.9 %
/l/	36.8%	93%
/m/	64.9%	86.5%
/n/	32.7%	95.9%
/tʃ'/	19.2%	59.2%
/r/	4%	16% , 52%
/s/	15.3%	39.8%
/ʂ/	34%	71%
/ʃ/	22%	48%
/ts/	5.6%	13.5%
/tʂ/	15%	29%
/x/	56.3%	56.3, no group
/f/	44.3%	44.3, no group
total	26%	53%, 56%

In light of phonetics, these 15 phonemes are divided into 7 groups: laminal /tʃ/ /tʃ'/ and /ʃ/;

retroflex /tʂ/ /tʂʰ/ and /ʂ/; apical /ts/ /tsʰ/ and /s/; nasal and lateral /m/ /n/ and /l/; vowel /a/ /o/ /i/ /y/ /u/ and /ʏ/; radical fricative /x/ and labio-dental fricative /f/ compose a group respectively. In table 1, the ‘accuracy over group’ refers to the accuracy of which a phoneme is classified into own group. /r/ is special, it is mixed with vowel group, especially with /a/ and /y/, so we count its classification accuracy according to /a/ /y/ /r/ group and vowel group respectively, they are 16% and 52% respectively. Table 1 shows that the classification accuracy of nasal and lateral group is the highest in all groups, and apical group is the lowest, perhaps because it is adjacent two groups, namely retroflex group and laminar group.

V. SUMMARY

In table 1 it is noticed that nasal and lateral have the best accuracy in all phoneme groups, because they have the smallest distance in SE space. /x/ and /f/ have also good accuracy, because they are apart from rest phonemes. Besides /x/, only 0.4% affricate and fricative are classified into vowel.

So, it was shown that SE parameter can express phonetic nature of each phoneme. In future work, the distribution of SE parameter will be investigated for obtaining better classification accuracy. We also take note of that SE value is dependent to signal sampling rate and SNR (Signal Noise Ratio), so we will find the transform using which SE value will make phoneme classification more robust to signal sampling rate variation and SNR variation.

REFERENCES

1. H. D.I. Abarbanel, *Analysis of Observed Chaotic Data*, Springer-Verlag, NewYork, 1996.
2. Michael T. Johnson, Richard J. Povinelli, Andrew C. Lindgren, Jinjin Ye, Xiaolin Liu, Kevin M. Indrebo, "Time-Domain Isolated Phoneme Classification using Reconstructed Phase Spaces," *IEEE Transactions on Speech and Audio Processing*, vol. **13**, no. 4, July, 458-466. 2005.
3. V. Pitsikalis, I. Kokkinos and P. Maragos, Nonlinear Analysis of Speech Signals: Generalized Dimensions and Lyapunov Exponents ,*Proc. European Conference on Speech Communication. & Technology (EUROSPEECH-2003)*, Geneva, Switzerland, Sep. 2003.
4. Kokkinos and P. Maragos, “Nonlinear Speech Analysis Using Models for Chaotic Systems”, *IEEE Transactions on Speech and Audio Processing*, Nov. 2005.
5. J. Ye, M. T. Johnson, and R. J. Povinelli, "Phoneme classification over the reconstructed phase space using principal component analysis," *ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP)*, Le Croisic, France, 11-16..
6. V. Pitsikalis and P. Maragos, “Speech analysis and feature extraction using chaotic models,” in *Proc. ICASSP, 2002*.

7. Michael T. Johnson, and Richard J. Povinelli, "Speech Recognition using Reconstructed Phase Space Features", *International Conference on Acoustics, Speech and Signal Processing 2003 (ICASSP03)*, Hong Kong, April 2003.
8. Xiaolin Liu, Richard J. Povinelli, and Michael T. Johnson, "Vowel classification by global dynamic modeling", *ISCA tutorial and research workshop on non-linear speech processing (NOLISP) 2003*, Le Croisic, France, May 2003
9. Kevin M. Indrebo, Richard J. Povinelli, Michael T. Johnson. "A Combined Sub-band and Reconstructed Phase Space Approach to Phoneme Classification," *ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP)*, Le Croisic, France, 107-110.(2003)