SPEECH ANALYSIS USING HIGUCHI FRACTAL DIMENSION

Jari Turunen, Tarmo Lipping & Juha T. Tanttu
Tampere University of Technology, Pori
P.O.Box 300, Pohjoisranta 11, FIN-28101 Pori
{jari.turunen, tarmo.lipping, juha.tanttu}@pori.tut.fi

Speech analysis for identification and recognition purposes is a demanding task, especially to find recognition parameters for some consonants. In this paper, we show the initial analysis results of speech corpus, sampled with two sampling rates, 22050 and 8000 Hz, using Higuchi fractal dimension. The study shows that it might be possible to use fractal dimension value for classification of for example plosives /k/, /p/ and /t/. However, the analysis seems to be very sensitive to the sampling frequency in speech analysis.

I. INTRODUCTION

Linear speech analysis and models has served successfully the speech processing areas for decades. However, there has also been an interest to research and evaluate nonlinear methods in order to find better way to model speech, especially for speech recognition purposes.

Several studies have found evidences for nonlinear behavior in speech. Different nonlinear techniques have been tested with time series over several decades in order to improve modeling and estimation when compared to linear methods. For example, the logarithmic a-law/μ-law compression in Pulse Code Modulation (PCM) coding has worked successfully over the years. However, the precise "practical" nonlinearity form for vocal tract model is not known and the search for good alternatives for the describing the human vocal tract with linear methods is currently going on. The Hammerstein model, Volterra series and Wiener filters have been tested experimentally as well as the chaotic time series modeling with very good results. However, the disadvantages are, when compared to the linear models, the more complex parameter computations and in some cases, the stability preservation. Also, different types of neural networks have been tested for several purposes in speech processing. Neural network is easy to design, train and test but there still remains a fear that the unique and documented experiment is unrepeatable even with the same data [1-17].

The nonlinear signal processing field is enormous, in that sense that the number of different functions, equations and/or systems that can be used for speech modeling and analysis is practically infinite.

The chaotic models have worked successfully for vowels and nasals [4, 5] and Teager energy operator [18 ,19] has been used to indicate several features of speech, for example speech resonances and modulations.

In this paper we study the effect of Higuchi Fractal dimension for different phonemes.

II. METHODS AND DATA

Higuchi fractal dimension [20] is a method developed for estimating the amount of self-similarity of the data. Higuchi [20] used his method for magnetic field data and in [21, 23-25] Higuchi's method was used for electroencephalography data.

Method described in [20], defines the discrete time series:

X[1], X[2], … ,X[n]

to be constructed to a new time series:

$$X_k^m; X[m], X[m+k], X[m+2k],..., X(m+[\frac{N-m}{k}]k),$$
$$m = 1,2,..., k$$

where m is the initial time and k is the interval time, N is the total number of samples. For example if k=3 and N=100, three time series are obtained as follows:

$$X_3^1; X(1), X(4), X(7),..., X(100)$$
$$X_3^2; X(2), X(5), X(8),..., X(98)$$
$$X_3^3; X(3), X(6), X(9),..., X(99)$$

The length of the curve, defined in [20] is:

$$L_m(k) = \frac{\left\{\left(\sum_{i=1}^{\left[\frac{N-m}{k}\right]} |X(m+ik) - X(m+(i-1)k|\right)\frac{N-1}{\left[\frac{N-m}{k}\right]k}\right\}}{k}$$

$L_m(k)$ represents normalized sum of "segment length". Each "segment length" represents the absolute value of difference between magnitude values of pair of points distant k samples, starting from m^{th} . sample. The length of the curve L(k) is mean of k values $L_m(k)$, for m=1,2,…k. The fractal dimension D is the least square estimate of the slope of the curve evaluated on $1..k_{max}$ values on L(k). If the curve is plotted on doubly logarithmic scales, for $1..k_{max}$ in ln(1/k) and L(k) in

Models and analysis of vocal emissions for biomedical applications. 4th international workshop. October 29-31, 2005 – Firenze, Italy.
Edited by C. Manfredi. ISBN 88-8453-320-1 (online)

ln(L(k)), the data should fall on a straight line with a slope (-D). It should be noted that Higuchi fractal dimension is not related with chaotic attractor dimension.

For example, the Higuchi value of white noise, with maximum amplitudes [-1,1] with k_{max} =10, is 1, and straight line with slope is zero.

The algorithm and evaluation was performed using Matlab environment. In our preliminary experiments, several k_{max} values were tested. If the k_{max} values was increased beyond 25 the Higuchi fractal dimension values tend to get closer to one as the k_{max} increases. Similarily, if the k_{max} value was decreased below 8, the Higuchi values tend to progress towards zero. The k_{max} value 10 gave best results in our experiments.

Table 1. Phonemes and their corresponding example words.

phon.	**word**	**phon.**	**word**	**phon**	**word**
/p/	pin	/tS/	chin	/i/	see
/b/	bay	/dZ/	jam	/a/	father
/t/	toy	/m/	me	/O/	sort
/d/	die	/n/	not	/Î/	bird
/k/	key	/N/	sing	/u/	too
/g/	get	/l/	light	/ei/	day
/f/	five	/r/	ring	/ai/	fly
/v/	van	/w/	win	/Oi/	boy
/T/	thick	/j/	yes	/ou/	go
/D/	then	/I/	sit	/au/	cow
/s/	see	/e/	get	/i«/	ear
/z/	zinc	/Q/	cat	/u«/	tour
/S/	ship	/Ã/	hut	/e«/	air
/Z/	measure	/A/	hot	/q/	[silence]
/h/	he	/U/	put	/«/	banana

Table 2. Number of different phonemes in each phoneme category.

phon.	**#**	**phon.**	**#**	**phon**	**#**
/p/	83	/tS/	89	/i/	154
/b/	61	/dZ/	52	/a/	96
/t/	191	/m/	72	/O/	162
/d/	66	/n/	236	/Î/	92
/k/	140	/N/	48	/u/	145
/g/	60	/l/	61	/ei/	156
/f/	211	/r/	152	/ai/	216
/v/	107	/w/	106	/Oi/	96
/T/	138	/j/	48	/ou/	151
/D/	61	/I/	153	/au/	95
/s/	205	/e/	173	/i«/	44
/z/	115	/Q/	72	/u«/	37
/S/	80	/Ã/	130	/e«/	68
/Z/	38	/A/	72	/q/	3
/h/	48	/U/	48	/«/	30

The OTAGO speech corpus [22] was used in the tests. The phonemes were manually checked and identified. The data, which was used in experiments, is presented in Table 1 and Table 2. The sampling frequency was 22050 Hz and total number of samples was 4661.

The phoneme lengths varied from minimum of 143 samples in /b/ to 10949 samples in /a/ sampled with 22050 Hz frequency. The /q/ is a silence "phoneme" recorded by the microphone (background noise). We preformed two tests with Higuchi analysis: the first one with 22050 sampling frequency and the second one with 8000 Hz sampling frequency. The lower sampling frequency was obtained by resampling the data from the original data by using Matlab "resample" command.

III. Results

The results are shown in Figures 1-6. The Figures 1-3 show the fractal dimension values for phonemes sampled at 22050 Hz and Figures 4-6 show the fractal dimension values for 8000 Hz data.

In the figures, the boxes show the lower quartile, median and the upper quartile of all sampled phoneme values. The lines show the deviation for the rest of phoneme data and outliers are presented with '+' sign.

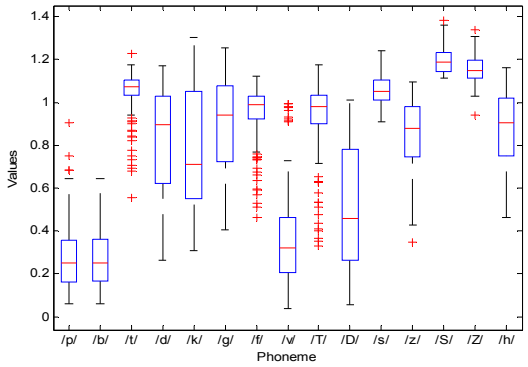

Figure 1. The fractal dimension values for phonemes /p/ to /h/ sampled at 22050 Hz frequency

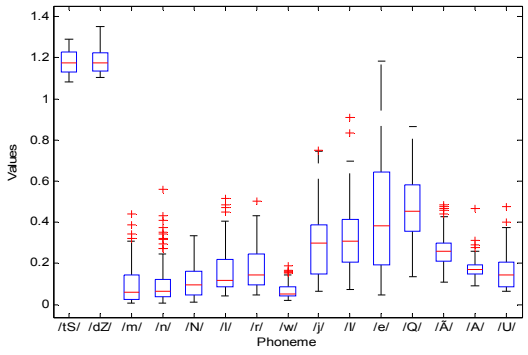

Figure 2. The fractal dimension values for phonemes /tS/ to /U/ sampled at 22050 Hz frequency

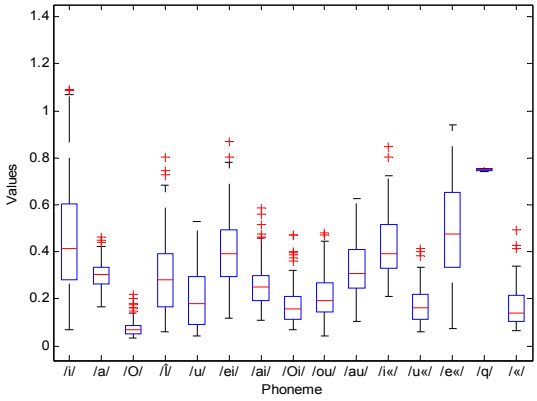

Figure 3. The fractal dimension values for phonemes /i/ to /«/ sampled at 22050 Hz frequency

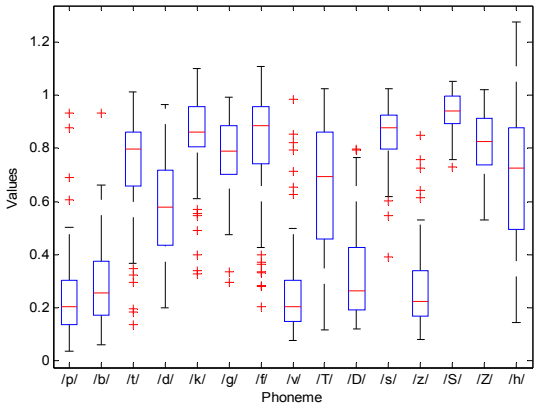

Figure 4. The fractal dimension values for phonemes /p/ to /h/ sampled at 8000 Hz frequency

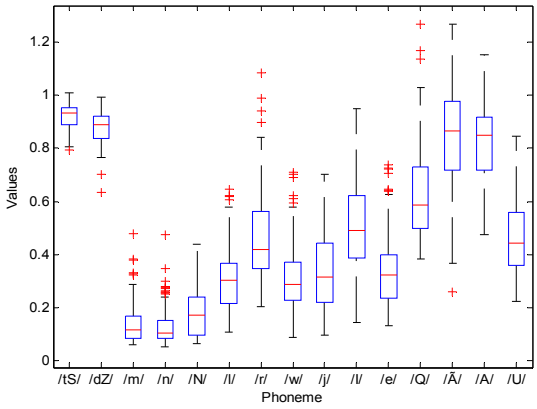

Figure 5. The fractal dimension values for phonemes /tS/ to /U/ sampled at 8000 Hz frequency

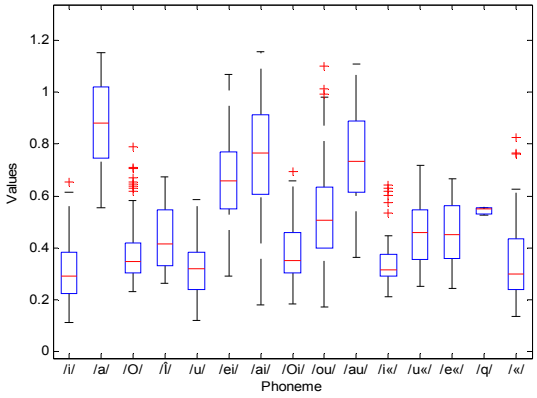

Figure 6. The fractal dimension values for phonemes /i/ to /«/ sampled at 8000 Hz frequency

IV. Discussion

When looking the Figures 1-3 with Fs=22050 Hz, the Higuchi Fractal dimension reveals interesting things from the phonemes. For example plosives /p/, /t/ and /k/ has been very difficult to separate from each other by using other methods, but is seems that they are possible to separate by using Higuchi fractal dimension. The quartile bars and the whole data deviation in plosive /k/ does not separate from /t/ as well as the /p/, although the median is different by visual inspection. In addition, the grouping of the fricatives can be seen from the figures 1-2. The fricatives /s/, /z/, /f/, /Ts/, /Tz/ and /q/ (that means background noise) medians are concentrated near one (from 0.8 to 1.2). This may also be interpreted that the Higuchi fractal dimension is very sensitive to background noise.

The nasals /n/ and /m/, semivowels /v/ and /w/ and all vowels have median values below 0.5. The Higuchi values of vowels, interestingly, seem to correlate with the vowel production mechanism somehow, in the sense of tongue position in the mouth. For example, lower fundamental frequency vowel /U/ has smaller Higuchi value when compared to higher fundamental frequency vowel /i/.

All seems to separate nicely with 22050 Hz recordings but unfortunately speech is usually sampled and transferred with much lower, 8000 Hz sampling frequency. The interesting phenomena do not appear in same depth anymore in lower sampling frequency. When looking the figures 4-6 the phonemes do not show similar behavior as they did in figures 1-3. The plosives /t/ and /p/ do separate in figure 4, but the quartile bars are much closer (and slightly overlapping) to each other than in figure 1. The values of phoneme /k/ are higher than in previous analysis. The fricatives /s/, /z/, /f/, /Ts/, /Tz/ and /q/ show overall dropping in median values especially in

the case of /z/. Also phonemes seem to have overall increase in the median values. The Higuchi values for vowels have higher values than in 22050 Hz sampling frequency, but in the case of vowel /i/, the fractal dimension values are dropped.

When thinking the Higuchi fractal dimension value computation, the differences between two sampling frequency results seems to follow the filtering properties. The Higuchi fractal dimension measures the amount of self-similarity by searching the difference between adjacent samples. The downsampling will smooth the fine structure of the signal.

The selection of k_{max}. parameter is also very important for the analysis. In our case the both sampling frequencies 22050 and 8000 Hz the k_{max}=10 seems to be good for speech analysis purposes. Several k_{max} values were tested, well beyond 50, because in vowels the fundamental cycle repetition is approximately 50-140 samples in 8000 Hz sampling frequency. k_{max} values beyond 50 provided fractal dimension values that are all approaching one rather than providing better separation between consonants and vowels. With lower sampling frequencies some critical information may be lost, which may be useful for recognition purposes with this method.

Higuchi fractal dimension is useful tool, and the algorithm is very simple and easy to compute providing single number for example analysis and recognition purposes.

REFERENCES

[1] Kubin G., "*Nonlinear Processing of Speech*", Speech Coding and Synthesis, Elsevier Science, Amsterdam, 1995.

[2] Townshend B., "*Nonlinear prediction of speech*", IEEE ICASSP: 425-428,1991.

[3] Langi A*., Soemintaputra K. & Kinsner W., "*Multifractal Processing of Speech Signals", IEEE ICICS: 527-531, 1997.

[4] Banbrook M., McLaughlin S. & Mann I., *"Speech Characterisation and Synthesis by Nonlinear Methods",* IEEE Trans. Speech and Audio Proc, 7 (1): 1-17,1999

[5] Miyano T., Nagami A., Tokuda I., Kazuyuki A., "Detecting nonlinear determinism in voiced sounds of Japanese vowel /a/", in *Int. Journal of Bifurcation and Chaos,* 10 (8), 1973-1979, 2000.

[6] Thyssen J., Nielsen H. & Hansen S., "*Non-linear short term prediction in speech coding*", IEEE ICASSP, (1): 185-188, 1994.

[7] Kumar A. & Gersho A., "*LD-CELP Speech Coding with Nonlinear Prediction*", IEEE Signal Processing Letters, 4 (4), 89-91,1997.

[8] *Ma N. & Wei G., "Speech Coding with Nonlinear Local Prediction Model*", IEEE ICASSP: 1101-1104, 1998.

[9] *Birgmeier M., Bernhard H. & Kubin G.,"Nonlinear Long-Term Prediction of Speech Signals*", IEEE ICASSP: 1283-1286, 1997.

[10] Kubin G., "*Synthesis and Coding of Continuous Speech with The Nonlinear Oscillator Model", IEEE ICASSP:* 267-270, 1996.

[11] Ohmura H. & Tanaka K., "*Speech Synthesis Using a Nonlinear energy Damping Model for The Vocal Folds Vibration Effect*", IEEE ICSLP, (2): Acoustic Analysis, #11, 1996.

[12] Abarbanel H., "*Chaotic Signals and Physical Systems*", IEEE ICASSP, (4): 113-116, 1992.

[13] Singer A., Wornell G. & Oppenheim A., "*Codebook Prediction: A Nonlinear Signal Modeling Paradigm*", IEEE ICASSP (5), 325-328, 1992.

[14] Díaz-de-Maria F. & Figueiras-Vidal A., "*Radial Basis Functions for Nonlinear Prediction of speech in Analysis-by-Synthesis Coders",* IEEE ICASSP: 788-791, 1995.

[15] Hennebert J., Hasler M. & Dedieu H., "*Neural networks in speech recognition",* Proc. 6th Microcomputer School, Prague, 1994.

[16] Fackrell J., "*Bispectral Analysis of Speech Signals*", doctoral dissertation, University of Edinburgh, 1996.

[17] Turunen J., Tanttu J., & Loula P., "*Hammerstein model for speech coding*", in *Eurasip JASP* (12)*,* 1238-1249, 2003.

[18] Teager H., "*Some Observations on Oral Air Flow During Phonation*", in *IEEE Transactions on ASSP*: 28, 599-601, 1980.

[19] Teager H., & Teager S., "Evidence ofor Nonlinear Production Mechanisms on Vocal Tract" in *NATO adv. Study Inst. On Speech Production and Speech Modelling,* Kluwer, Bonas France,1990.

[20] Higuchi T., "Approach to an irregular time series on the basis of the fractal theory", in *Physica D*, 31, 277-283, 1988.

[21] Accardo A., Affinito M., Carrozzi M. & Bouquet F., "*Use of fractal dimension for the analysis of electroencephalographic time series*", in *Biological Cybernetics*, 77 (5), 339-350, 1997.

[22] OTAGO speech corpus, available URL: http://translator.kedri.info/datasets/corpus/otago.

[23] T. Lipping, E. Olejarczyk and M. Parts. Fractal dimension analysis of the effects of photic and microwave stimulation on the brain function. Proceeedings of the World Congress on Biomedical Engineering and Medical Physics, Sydney, Australia, 24-29.08.2003 (on CD ROM).

[24] T. Lipping, E. Olejarczyk and M. Parts. Analysis of photo-stimulation and microwave stimulation effects on EEG signal using Higuchi's fractal dimension method. In *Optical Methods, Sensors, Image Processing, and Visualization in Medicin,* A. Nowakowski and B. B. Kosmowski eds., Proc. SPIE, vol. 5505 (SPIE, Bellingham, WA, 2004), pp. 174-178.

[25] A. Anier, T. Lipping, S. Melto, S. Hovilehto. Higuchi fractal dimension and spectral entropy as measures of depth of sedation in intensive care unit. Proc of the 26th IEEE EMBS Annual International Conference, San Fracisco, USA, September 1-5, 2004, pp. 526-529.