Jurnal
Teknologi

Full Paper

LPC AND ITS DERIVATIVES FOR STUTTERED SPEECH RECOGNITION

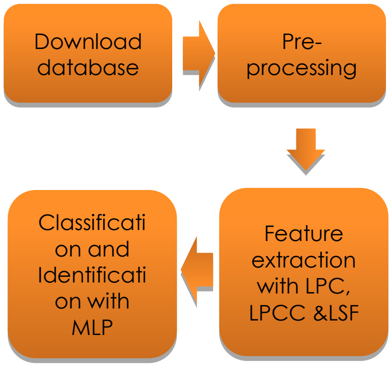
Sabur Ajibola Alim*, Nahrul Khair Alang Rashid, Wahju Sediono, Nik Nur Wahidah Nik Hashim

Mechatronics Engineering Department, Kulliyyah of Engineering, International Islamic University Malaysia, Malaysia

Article history
Received
15 May 2015
Received in revised form
1 July 2015
Accepted
11 August 2015

*Corresponding author
moaj1st@yahoo.com

Graphical abstract


Download database
Pre-processing
Feature extraction with LPC, LPCC &LSF
Classification and Identification with MLP

Abstract

Stuttering or stammering is disruptions in the normal flow of speech by dysfluencies, which can be repetitions or prolongations of phoneme or syllable. Stuttering cannot be permanently cured, though it may go into remission or stutterers can learn to shape their speech into fluent speech with an appropriate speech pathology treatment. Linear Prediction Coefficient (LPC), Linear Prediction Cepstral Coefficient (LPCC) and Line Spectral Frequency (LSF) were used for the feature extraction, while Multilayer Perceptron (MLP) was used as the classifier. The samples used were obtained from UCLASS (University College London Archive of Stuttered Speech) release 1. The LPCC-MLP system had the highest overall sensitivity, precision and the lowest overall misclassification rate. LPCC-MLP system had challenges with F3, the sensitivity of the system to F3 was negligible, similarly, the precision was moderate and the misclassification rate was negligible, but above 10%.

Keywords: Stuttering, linear prediction coefficient, linear prediction cepstral coefficient, line spectral frequency

1.0 INTRODUCTION

Only about 5-10% of the entire human population has a perfectly normal form of oral communication in relation to numerous speech features and healthy voice; and the rest of the population of about 90-95% exhibit one form of the disorder or the other such as stuttering, apraxia of speech, dysarthria and cluttering [1]. Stuttering or stammering can be defined as a unintentional disruption in the normal flow of speech by dysfluencies, which include repetitive pronunciation, prolonged pronunciation, blocked or stalled pronunciation at the phoneme or the syllable level [2]–[4].

One of the usual features of stuttering is in its variability, and that it may be manipulated and influenced by a wide variety of strategies [5]. Stuttering cannot be permanently cured; it may go into remission for a time, or a stutterer can learn to shape their speech into fluent speech with the appropriate speech pathology treatment. This shaping has its effects on the tempo, loudness, effort, or duration of the utterances [3], [6]. Nearly 2% of adults exhibit stuttering, while about 5% of children stutter [7], [8]. The stuttering that is prevalent in children is called developmental stuttering [8].

Low speech recognition rate is the bottleneck that impedes effective detection of stuttered speech [4]. Some previous research studies on recognition of repetition and prolongation in stuttered speech include the use of Linear Prediction Cepstral Coefficient (LPCC) for feature extraction, while using k-Nearest Neighbor (k-NN) & Linear Discriminant Analysis (LDA) as classifiers [9], Mel Frequency Cepstral Coefficient (MFCC) for feature extraction, while using k-NN & LDA as classifiers [2], MFCC for feature extraction, while using Support Vector Machines (SVM) as classifier

[10] and MFCC for feature extraction, while using Dynamic Time Warping (DTW) as classifier [11]. Stuttered speech is rich in dysfluencies, which are responsible for lower Automatic Speaker Recognition (ASR) rates. This research looks into the use of LPC and its derivatives, LPCC and LSF for feature extraction, while Multilayer Perceptron was used as the classifier in each of the cases considered.

2.0 EXPERIMENTAL

2.1 Linear Prediction Coefficient (LPC)

Linear Prediction Coefficients (LPC) models the human vocal tract [12] and gives good speech feature estimation. It analyzes the speech signal by estimating the formants and eliminating their effects from the speech signal, followed by the estimation of the intensity and frequency of the remaining buzz. Its solution is a difference equation, which shows all the samples of the signal as a linear combination of previous samples, an equation called a linear predictor. The coefficients of the difference equation (the prediction coefficients) characterize the formants, and thus the LPC system estimates these coefficients [13]. Other features that can be extracted from LPC include Reflection Coefficient (RC), Linear Predication Cepstral Coefficients (LPCC), Log Area Ratio (LAR), Arcus Sine Coefficients (ARCSIN) and Line Spectral Frequency (LSF) [14].

2.2 Linear Prediction Cepstral Coefficient (LPCC)

The Linear Prediction Cepstral Coefficients (LPCC) are linear prediction-derived cepstral coefficients. They are derived from LPC computed spectral envelope [15] and are standardized between +1 and -1 [16], [17]. The LPC based cepstral coefficients are the coefficients of the Fourier transform representation of the logarithmic magnitude spectrum [18], [19] of the LPC. In general, one of the most attractive features of the cepstrum which makes it a good candidate for usage in speaker recognition is its inherent invariance toward linear spectral distortions [20]. LPCC utilizes an all-pole filter to model the human vocal tract with speech formants captured by the poles of the all-pole filter. The narrow band (up to 4 KHz) of LPCC features works well in a clean environment. However, the linear predictive spectral envelope shows large spectral distortion in noisy environments, resulting in significant performance degradation [21].

2.3 Line Spectral Frequency (LSF)

Line Spectral Frequency (LSF) exhibits ordering and distortion independent properties. These properties enable the representation of the high frequencies associated with less energy using fewer bits [22]. LSF is an alternative to the direct form predictor coefficients or the lattice form reflection coefficients for representing the filter response. The direct form coefficient representation of the LPC filters is not conducive for efficient quantization. Instead, nonlinear functions of the reflection coefficients are often used as transmission parameters. These parameters are preferable because they have a relatively low spectral sensitivity [23]. It has been found that the line spectral frequency (LSF) representation of the predictor is particularly well suited for quantization and interpolation. Theoretically, this is motivated by the fact that the sensitivity matrix relating the LSF-domain squared quantization error to the perceptually relevant log spectrum is diagonal [24].

2.4 Multilayer Perceptron (MLP)

Multilayer perceptron (MLP) is one of many different types of existing neural networks. It comprises of a number of neurons connected together to form a network. This network has three layers which are input layer, one or more hidden layer(s) and output layer with each layer containing multiple neurons [14]. A neural network is able to classify the different aspects of the behaviors of a system, knows what is going on at the instant, diagnoses whether it is correct or faulty, forecasts what it will do next, and if required responds with what do next [25], [26]. These neuron connections are in forward direction only.

2.5 Methodology

The stuttered speech database used in this research is the UCLASS (University College London Archive of Stuttered Speech) release 1. These speech recordings were collected at University College London (UCL) for a number of years. The recordings were from people (mostly children) who were referred to clinics in London for assessment of stuttering. Release One recordings were all monologs and from speakers with a wide range of ages [27]. All the samples were quantized at a bit rate of 16 bits. Table 1 below shows the age, sex and sampling frequency of the 8 samples used for this experiment. Each sample was divided into smaller bits of 10 seconds and 11 samples per sample.

The relevant features were extracted from each sample, using Linear Prediction Coefficient (LPC), Linear Prediction Cepstral Coefficient (LPCC) and Line Spectral Frequency (LSF). A three layer multilayer perceptron with 215 hidden neurons was used for the classification and identification. The confusion matrix plot of the designed system was plotted. From the confusion matrix plot, the sensitivity, precision and the misclassification rate were computed as performance measures. Sensitivity is the measure of the correctly identified samples, while precision is the measure of the ability of the system to reproduce the same output for the same set of input and misclassification rate is the measure of the percentage of the incorrectly identified samples to the total number of samples.

Table 1 Summary of the samples used

	Age	**Sex**		**Sampling frequency**
1	15y2m	F	F1	44.1 kHz
2	17y2m	F	F2	44.1 kHz
3	15y11m	F	F3	44.1 kHz
4	12y11m	F	F4	44.1 kHz
5	16y4m	M	M1	22.05 kHz
6	17y9m	M	M2	44.1 kHz
7	19y5m	M	M3	44.1 kHz
8	16y9m	M	M4	22.05 kHz
mean	16y5m			

3.0 RESULTS AND DISCUSSION

The tool that was used to understand and interpret the results that were obtained is the classification developed by Best in 1981. This classification can be used to describe the significance of probability of any experiment. It is listed as follows: 0 - 0.20 (0 - 20%) – negligible, 0.20 - 0.40 (20 - 40%) – low, 0.40 - 0.60 (40 - 60%) – moderate, 0.60 - 0.80 (60 - 80%) – substantial & 0.80 - 1.00 (80 - 100%) – high.

3.1 Sensitivity

Table 2 shows the sensitivity of LPC, LPCC and LSF feature extractor in conjunction with MLP as the classifier. The table shows that the sensitivity of all the samples considered for LPCC except for F3 all had between 80 and 100% which is categorized as high. Although, it is desirable that the sensitivity of the algorithm should be between 90 and 100%, however, 80-90% can also be accepted based on the categorization of Best (1981). Similar to LPCC, LPC's sensitivity to F3 was the least and categorized as negligible as it falls below 20%. Also, F1, F2, F4 & M3 all had between 60 and 80%, which is categorized as substantial. Furthermore, in the case of LSF, only F1, F2 & M4 falls into the category high, while F4 & M1 fall into the category low (20-40%).

In the design of a speaker verification system, it is desirable for the system to be able to efficiently and effectively sense and detect the samples that are used to train it. As such, it is expected that the sensitivity of the designed system to each of the samples is 100%. The LPCC-MLP system had a high sensitivity for most of the training samples except for F3.

Table 2 Sensitivity

	LPC (%)	LPCC (%)	LSF (%)
F1	63.63	100	90.91
F2	72.72	90.91	100
F3	18.18	18.19	72.73
F4	72.72	81.82	36.36
M1	100	100	27.27
M2	100	100	72.73
M3	72.72	100	72.73
M4	100	100	100

3.2 Precision

Table 3 shows how precise the systems designed with LPC, LPCC and LSF in conjunction with MLP classifier was in its ability to repeatedly achieve high classification. The precision of LPC for F2, M1, M2, M3 & M4 all fall into the category high (80-100%), while F1 is between 60 & 80%, substantial, F4 is between 40 & 60%, moderate and F3 is between 20 & 40%, low. In addition, for LPCC, all were 100%, high, except for F3 & F4 which fall between 40 & 60%, Moderate. Furthermore, for LSF, f1, f2, m3 & m4 all had 100% precision, which is categorized as high, while f3, f4 & m2 all fall into the category, moderate (40-60%), and m1 fall in between 20 & 40%, low.

Speaker verification systems are repeatedly used systems, as a result, the ability of such a system to repeatedly be able to identify all the samples used to train the system is important. LPCC-MLP system has a high precision for all the samples used to train it except F3 and F4.

Table 3 Precision

	LPC (%)	LPCC (%)	LSF (%)
F1	70	100	100
F2	80	100	100
F3	28.57	40	44.44
F4	44.44	50	57.14
M1	100	100	33.33
M2	91.7	100	57.14
M3	88.89	100	100
M4	100	100	100

3.3 Misclassification Rate

Table 4 shows the misclassification rate, the probability that a sample would be wrongly classified. For a system to be highly reliable, the misclassification rate has to be below 10%, however, negligible, (0-20%) would be appropriate based on the categorization being used. In line with this, all the misclassification rates obtained were negligible. The system designed with LSF had a below 10% misclassification rate for F1, F2, M3 & M4, while the other four are between 10 and 20%. Furthermore, for LPCC, only F3 & F4 had misclassification rates between 10 and 20%, while the others were below 10%. Similar to LPCC, LPC had a misclassification rate of above 10% for only F3 & F4, the others were below 10%.

Correctly identifying each of the samples used to train the system is an important characteristic of speaker identity system. The lower the misclassification rate, the better the system, therefore, it is desirable for it to be 0%. For the LPCC-MLP system has the overall misclassification rates. Samples F3 and F4 have misclassification rates that is more than 5%.

Table 4 Misclassification Rate

	LPC (%)	LPCC (%)	LSF (%)
F1	8	0	1.14
F2	5.7	1.14	0
F3	16	13.64	14.77
F4	14.77	12.5	11.36
M1	0	0	15.91
M2	1.14	0	10.47
M3	4.45	0	3.41
M4	0	0	0

4.0 CONCLUSION

In conclusion, three sets of systems were designed and tested, LPC-MLP, LPCC-MLP and LSF-MLP. They all showed great potentials, however, the LPCC-MLP system had the highest overall sensitivity, precision and the lowest overall misclassification rate. This was closely followed by LPC in overall sensitivity, precision and misclassification rate. And lastly the LSF had the least results obtained. In the case of the LPCC, the system had challenges with F3, the sensitivity of the system to F3 was negligible, similarly, the precision was moderate and the misclassification rate was negligible, but above 10%.

References

[1] G. Manjula and M. Kumar. 2014. Stuttered Speech Recognition for Robotic Control Work. 3(12).

[2] L. Chee, O. Ai, M. Hariaran, and S. Yaacob. 2009. MFCC based Recognition of Repetitions and Prolongations in Stuttered Speech using k-NN and LDA. In *SCOReD2009-Proceedings of 2009 IEEE Student Conference on Research and Development.*

[3] M. Hariharan, L. S. Chee, and S. Yaacob. 2012. Analysis of Infant Cry Through Weighted Linear Prediction Cepstral Coefficients and Probabilistic Neural Network. *J. Med. Syst.* 36(3): 1309-15.

[4] J. Zhang, B. Dong, and Y. Yan. 2013. A Computer-Assist Algorithm to Detect Repetitive Stuttering Automatically. In *2013 International Conference on Asian Language Processing (IALP)*. 249-252.

[5] T. Voigt, K. Hewage, and P. Alm. 2014. Smartphone Support for Persons Who Stutter. In *13th International Symposium on Information Processing In Sensor Networks*. 293-294.

[6] S. Awad. 1997. The Application of Digital Speech Processing to Stuttering Therapy. In *Instrumentation and Measurement Technology Conference*. 1361-1367.

[7] E. G. Conture and J. S. Yaruss. 2002. Treatment Efficacy Summary. *Am. speech-language Hear. Assoc.* 1993: 20850.

[8] C. Oliveira, D. Cunha, and A. Santos. 2013. Risk Factors for Stuttering in Disfluent Children with Familial Recurrence. *Audiol. Res.* 18(1): 43-49.

[9] L. S. Chee, O. C. Ai, M. Hariharan, and S. Yaacob. 2009. Automatic detection of prolongations and repetitions using LPCC. In *International Conference for Technical Postgraduates 2009. TECHPOS 2009*. 1-4.

[10] J. Pálfy and J. Pospíchal. 2011. Recognition of Repetitions Using Support Vector Machines. *Signal Process. Algorithms*.

[11] K. M. Ravikumar, B. Reddy, R. Rajagopal, and H. C. Nagaraj. 2008. Automatic Detection of Syllable Repetition in Read Speech for Objective Assessment of Stuttered Disfluencies. In *Proceedings of World Academy Science, Engineering and Technology*. 270-273.

[12] K. T. Al-Sarayreh, R. E. Al-Qutaish, and B. M. Al-Kasasbeh. 2009. Using the Sound Recognition Techniques to Reduce the Electricity Consumption in Highways. *J. Am. Sci.* 5(2): 1-12.

[13] S. Agrawal, A. K. Shruti, and C. R. Krishna. 2010. Prosodic Feature Based Text Dependent Speaker Recognition Using Machine Learning Algorithms. *Int. J. Eng. Sci. Technol*. 2(10): 5150-5157.

[14] R. Kumar, R. Ranjan, S. K. Singh, R. Kala, A. Shukla, and R. Tiwari. 2009. Multilingual Speaker Recognition Using Neural Network. *Proc. Front. Res. Speech Music. FRSM*. 1-8.

[15] K. M. Ravikumar, R. Rajagopal, and H. C. Nagaraj, 2009. An Approach for Objective Assessment of Stuttered Speech Using MFCC Features. *ICGST Int. J. Digit. Signal Process. DSP*. 9(1): 19-24.

[16] S. Ismail and A. bin Ahmad. 2004. Recurrent Neural Network with Backpropagation Through Time Algorithm for Arabic Recognition. *Proc. 18th ESM Magdeburg, Ger*. 13-16.

[17] A. M. Ahmad, S. Ismail, and D. F. Samaon. 2004. Recurrent Neural Network with Backpropagation Through Time for Speech Recognition. In *Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on*. 1: 98-102.

[18] M. M. El Choubassi, H. E. El Khoury, C. E. J. Alagha, J. a. Skaf, and M. a. Al-Alaoui. 2003. Arabic Speech Recognition Using Recurrent Neural Networks. In *Proceedings of the 3rd IEEE International Symposium on Signal Processing and Information Technology (IEEE Cat. No.03EX795)*. 543-547.

[19] Q.-Z. Wu, I. Chang Jou, and S.-Y. Lee. 1997. On-line Signature Verification using LPC cepstrum and Neural Networks. Syst. Man, Cybern. Part B Cybern. *IEEE Trans*. 27(1): 148-153.

[20] H. Beigi. 2011. *Fundamentals of Speaker Recognition*.

[21] Q. Li. 2012. *Speaker Authentication*. Springer-Verlag Berlin Heidelberg.

[22] V. Namburu. 2001. Speech Coder Using Line Spectral Frequencies of Cascaded Second Order Predictors.

[23] P. Kabal and R. Ramachandran. 1986. The Computation of Line Spectral Frequencies using Chebyshev polynomials. *Acoust. Speech Signal*.

[24] W. B. Kleijn, T. Bäckström, and P. Alku. 2003. On Line Spectral Frequencies. *IEEE Signal Process. Lett*. 10(3): 75-77.

[25] M. A. Al-Alaoui, L. Al-Kanj, J. Azar, and E. Yaacoub. 2008. Speech Recognition Using Artificial Neural Networks and Hidden Markov Models. IEEE Multidiscip. *Eng. Educ. Mag*. 3(3): 77-86.

[26] S. S. Haykin. 2009. *Neural Networks and Learning Machines*. vol. 3. Pearson Education Upper Saddle River.

[27] P. Howell, S. Davis, and J. Bartrip. 2009. The University College London Archive of Stuttered Speech (UCLASS). *J. Speech, Lang. Hear. Res*. 52(2): 556-569.