Classification of Fricatives Using Novel Modulation Spectrogram Based Features

Kewal D. Malde, Anshu Chittora, and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology,
Gandhinagar, Gujarat, India
{kewal_dhiraj,anshu_chittora,hemant_patil}@daiict.ac.in

Abstract. In this paper, we propose the use of a novel feature set, *i.e., modulation spectrogram* for *fricative classification*. Modulation spectrogram gives *2*-dimensional (*i.e.*, *2-D*) feature vector for each phoneme. Higher Order Singular Value Decomposition (HOSVD) is used to reduce the size of large dimensional feature vector obtained by modulation spectrogram. These features are then used to classify the fricatives in five broad classes on the basis of place of articulation (*viz., labiodental*, *dental*, *alveolar*, *post-alveolar* and *glottal*). Four-fold cross-validation experiments have been conducted on TIMIT database. Our experimental results show *89.09 %* and *87.51 %* accuracies for recognition of place of articulation of fricatives and phoneme-level fricative classification, respectively, using *3-nearest neighbor classifier*.

Keywords: Fricative classification, modulation spectrogram, HOSVD, place of articulation, acoustic frequency and modulation frequency.

1 Introduction

The main purpose of this paper is to use modulation spectrogram-based features to distinguish between various fricatives sounds extracted manually from TIMIT database. Consonants are classified in three broad classes based on the manner of articulation, *viz.*, stops, affricates and fricatives. Fricatives class is the largest class of consonants in TIMIT database. In total, there are *10* fricatives divided in five broad classes based on *place* of articulation, *viz.*, labiodental (*/f/*, */v/*), dental (*/th/*, */dh/*), alveolar (*/s/*, */z/*), post-alveolar (*/sh/*, */zh/*) and glottal (*/hh/*, */hv/*). Fricatives can also be classified as *unvoiced* and *voiced* on the basis of voicing activity, *i.e.*, relaxed or vibrating vocal folds. From all the examples of the fricatives mentioned above in pairs, the former in each pair falls under unvoiced fricatives while later falls under voiced fricatives [1-4].

Fricatives are generated by a narrow constriction in the vocal tract, giving rise to a steady *frication noise*. The degree of constriction also plays an important role in spectral characteristics (as secondary effect with vocal tract being the primary one). Place of articulation and the spectral characteristics are influenced on the basis of location of constriction by the tongue (at the back, center, or front of the oral tract, as well as at the teeth or lips) [1].

P. Maji et al. (Eds.): PReMI 2013, LNCS 8251, pp. 134–139, 2013.

Prior work on fricative classification on TIMIT database is based on the knowledge-based, acoustic-phonetic features where Seneff's auditory model is used as front-end processing. An accuracy of *87 %* was obtained for the overall classification of fricatives (on training data set) [5-7].

In this paper, we introduce the use of *2-D* representation of modulation spectrogram based for classification of fricatives. Modulation spectrogram-based feature set integrates the concept of sensory perception with signal processing methodology to achieve a significant improvement in the representation and coding of acoustic signals [8-9]. Recently, modulation spectrogram has been used in voice quality classification, pathology classification and speech recognition [10].

2 Modulation Spectrogram

Modulation spectrogram is the visual representation of the spectrum of the combination of *acoustic* and *modulation* frequencies in a speech segment. The spectrogram is square of the magnitude of short-time Fourier transform (STFT) of speech signal, representing the energy distribution of the signal *w.r.t. time* and *acoustic* frequency parameters. Modulation spectrogram is the spectrum of magnitude spectra of speech segment, representing energy distribution of the signal w.r.t. *acoustic* and *modulation* frequency parameters. Each phoneme has different combination of *acoustic* and *modulation* frequencies (depending on the constriction formed by the tongue). This motivated us to use modulation spectrogram for fricative classification. Modulation spectrogram is given by [10],

$$X_m(k) = \sum_{n=-\infty}^{\infty} h(mM - n)x(n)W_{I_A}^{kn}, \tag{1}$$

$$X_l(k,i) = \sum_{m=-\infty}^{\infty} g(lL - m) \mid X_m(k) \mid W_{I_M}^{im}, \tag{2}$$

where $k = 0, 1, ..., (I_A\text{-}1)$ is *acoustic* frequency, $W_{I_A} = e^{-j(2\pi/I_A)}$, $W_{I_M} = e^{-j(2\pi/I_M)}$, $i = 0, 1, ..., (I_M\text{-}1)$ is *modulation* frequency; n is the sample index in time-domain, m is the frame index in equation *(1)* and l is the frame index in equation *(2)*; $h(n)$ and $g(m)$ are the analysis window for acoustic and modulation frequency with hop sizes M and L, respectively. Here, $|X_l(k,i)|^2$ represents the modulation spectrogram of signal $x(n)$.

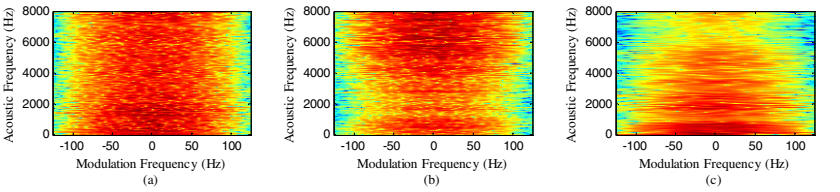

Fig. 1. Modulation Spectrogram: (a) labiodental */f/*, (b) alveolar */s/*, (c) glottal */hh/*

From some of the examples of modulation spectrogram shown in Fig. 1, we can observe that phonemes from different classes have their energy distributed across different regions of acoustic frequency (f_a) and modulation frequency (f_m) (as the mode of constriction formed by the tongue is different for each phoneme); hence, it can serve as a potential *cue* for classification of fricative sounds.

3 Higher Order Singular Value Decomposition (HOSVD)

Higher Order Singular Value Decomposition (HOSVD) is applied to *3-D* feature set. HOSVD theorem is used for dimension reduction of a *3-D* tensor as described in [11]. First of all, modulation spectrogram which is a *2-D* feature vector is obtained for each phoneme segment. Then a *3-D* tensor is formed by *stacking 2-D* modulation spectrogram feature vector (*i.e.*, $B \in \mathbb{R}^{I_A \times I_M}$ by considering only positive modulation frequencies taking symmetric nature of modulation spectrogram into consideration) for all phoneme samples (say I_S) under each class. Thus, the tensor $A \in \mathbb{R}^{I_A \times I_M \times I_S}$ can be represented in HOSVD form as:

$$A = S \times_1 U_A \times_2 U_M \times_3 U_S, \tag{3}$$

where S is the core tensor with same dimension as A. $U_A \in \mathbb{R}^{I_A \times I_A}$, $U_M \in \mathbb{R}^{I_M \times I_M}$ and $U_S \in \mathbb{R}^{I_S \times I_S}$ are the *unitary* matrices of the corresponding subspaces of I_A, I_M and I_S. The matrices U_A and U_M can be obtained by *unfolding* tensor A. U_A and U_M are obtained from SVD representation of the unfolded matrices $A_A \in \mathbb{R}^{I_M \times I_A I_S}$ and $A_M \in \mathbb{R}^{I_A \times I_M I_S}$. Then, $\hat{U}_A \in \mathbb{R}^{I_A \times R_A}$ and $\hat{U}_M \in \mathbb{R}^{I_M \times R_M}$ are obtained from these unitary matrices U_A and U_M, by retaining only first R_A and R_M vectors, respectively. Now, using these obtained $\hat{U}_A$ and $\hat{U}_M$, we get the reduced tensor as,

$$\hat{A} = S \times_1 \hat{U}_A \times_2 \hat{U}_M \times_3 U_S. \tag{4}$$

And finally un-stacking the reduced tensor $\hat{A}$, reduced feature vector $K \in \mathbb{R}^{R_A \times R_M}$ (per phoneme) is obtained as,

$$K = B \times_1 \hat{U}_A^T \times_2 \hat{U}_M^T = \hat{U}_A^T \cdot B \cdot \hat{U}_M. \tag{5}$$

4 Experimental Setup

4.1 Database Used

TIMIT database contains a total of *6300* sentences, *10* sentences spoken by each of *630* speakers from *8* major dialect regions of the United States. In addition to a speech waveform (*i.e.*, **.wav*) file, TIMIT corpus includes three associated transcription files (*viz.*, **.txt*, **.wrd*, **.phn*). With the help of phonetic transcription (*i.e.*, **.phn*) file we

obtained fricative segments. Out of all obtained fricative segments, we randomly selected *1000* samples of each phoneme (from both training and testing dataset) under *fricative* class. In our work, we have considered all *10 fricatives.*

4.2 Feature Extraction

First of all, find the modulation spectrogram for each phoneme using COG (center of gravity) method [12]. In modulation spectrogram, Y-axis represents *acoustic* frequency which ranges from 0 to $F_s/2$ (here, $F_s = 16$ *kHz*) (equally spaced in *54* bins) and X-axis represents *modulation* frequency which ranges from *-100* Hz to *+100*Hz (equally spaced bins). The modulation frequency dimension varies with respect to phoneme class and phoneme duration. For making the classification task easier, f_m dimension is fixed to *140* by *zero padding*. Since the modulation spectrogram is symmetric around $f_m = 0$, only positive modulation frequencies is considered for analysis. Using the *symmetry* property of modulation spectrogram the feature size is reduced from 54*x*140 to 54*x*71. To further reduce the feature dimension, HOSVD theorem is applied which gives a feature vector of dimension $f_a \times f_m$.

5 Experimental Results

Four-fold cross-validation experiments have been conducted on *56* different combinations of f_a and f_m. We have considered following values of f_a and f_m , *viz.*, f_a = [3 4 5 6 7 8 9] and f_m = [3 4 5 6 7 8 9 12]. In order to quote statistical significance of our experimental results, complete phoneme set is *randomly* divided into training and testing datasets (as per training-testing ratio), for every experiment.

Results of *4*-fold cross-validation of recognition of place of articulation of fricatives and phoneme-level fricative classification with feature vector dimension of 3×3 and *75:25 %* training-testing ratios are shown in Table 1 and Table 2, respectively. Fig. 2 shows the effect on classification accuracy due to change in acoustic (f_a) and modulation (f_m) frequencies. From Fig. 2, we can see that initially accuracy increases as the value f_a and f_m is increased (as the phoneme are more closely related in spectral characteristics, we need relatively higher dimension of feature vector for better classification of data), however, after certain amount of increase in the values of f_a and f_m , any further increase in the value decreases the classification accuracy (as it introduces more amount of redundancies). It can also be observed that for lower value of f_m (i.e., $f_m = 3$ and $f_m = 4$) accuracy is *almost constant*. We can also infer that change in the value of f_m affects the accuracy more significantly as compared to the change in the value of f_a. Best classification accuracy is obtained for feature of dimension 3×8 . However, we have selected the optimum feature vector dimension as 3×3 since the improvement in the accuracy is approximately *1 %* while the dimension is reduced by a factor of approximately *2.67*. The reduction in feature vector dimension is advantageous in reducing computational *complexity* and computation *time*.

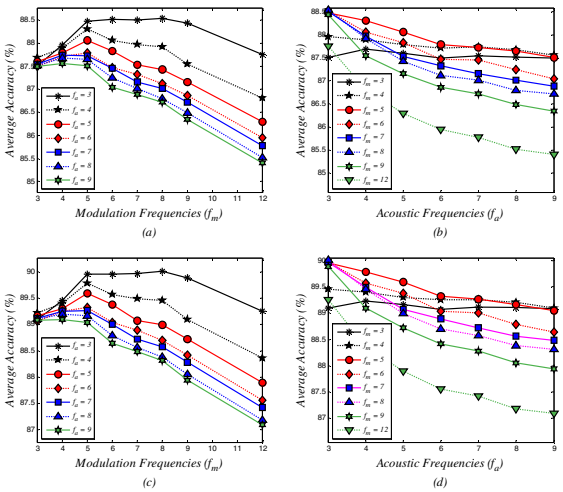

Fig. 2. Average accuracy (in %) for fricative consonant classification at phoneme-level: Effect on classification accuracy due to change in (a) modulation frequency (f_m), (b) acoustic frequency (f_a); Average accuracy (in %) for recognition of place of articulation of fricatives: Effect on recognition accuracy due to change in (c) modulation frequency (f_m), (d) acoustic frequency (f_a). In all sub-figures, units of acoustic and modulation frequencies are in bins.

Table 1. Confusion matrix for accuracy (in %) for recognition of place of articulation of fricatives. Average accuracy is *89.09 %*.

			Detected				
			Labiodental	Dental	Alveolar	Post-alveolar	Glottal
			/f/, /v/	*/th/, /dh/*	*/s/, /z/*	*/sh/, /zh/*	*/hh/, /hv/*
Actual	Labiodental	*/f/, /v/*	**88.22**	3.43	2.97	1.97	3.41
	Dental	*/th/, /dh/*	3.45	**89.85**	2.34	1.41	2.94
	Alveolar	*/s/, /z/*	2.92	2.34	**88.27**	3.44	3.03
	Post-alveolar	*/sh/, /zh/*	2.01	1.37	3.48	**90.81**	2.33
	Glottal	*/hh/, /hv/*	3.43	2.97	2.98	2.32	**88.30**

Table 2. Confusion matrix for phoneme-level fricative classification accuracy (in %). Average accuracy is *87.51 %*.

		Detected									
		/f/	*/v/*	*/th/*	*/dh/*	*/s/*	*/z/*	*/sh/*	*/zh/*	*/hh/*	*/hv/*
Actual	*/f/*	**85.89**	1.64	2.71	1.13	1.67	1.75	1.50	0.57	1.91	1.23
	/v/	1.61	**87.30**	1.65	1.37	1.22	1.29	1.29	0.58	1.78	1.90
	/th/	2.73	1.68	**85.70**	1.30	1.63	1.63	1.36	0.44	2.02	1.51
	/dh/	1.08	1.42	1.32	**91.37**	0.70	0.73	0.75	0.27	1.26	1.10
	/s/	1.63	1.24	1.55	0.71	**86.21**	2.10	2.66	1.01	1.44	1.45
	/z/	1.67	1.30	1.64	0.79	2.08	**86.15**	2.28	0.94	1.48	1.68
	/sh/	1.53	1.37	1.37	0.70	2.71	2.24	**85.39**	1.24	1.49	1.95
	/zh/	0.58	0.55	0.43	0.23	0.98	1.03	1.21	**93.77**	0.52	0.71
	/hh/	1.93	1.78	2.11	1.24	1.39	1.50	1.44	0.54	**86.48**	1.59
	/hv/	1.28	1.87	1.43	1.16	1.41	1.67	1.96	0.70	1.73	**86.80**

6 Summary and Conclusions

In this paper, we have proposed the use of modulation spectrogram-based features using a simple classifier, *i.e.*, *3*-nearest neighbor classifier, for fricative classification. Modulation spectrogram gives good classification accuracy for recognition of place of articulation of fricatives and phoneme-level fricative classification. Feature vector dimension of 3×3 after applying HOSVD theorem is considered as optimum. Modulation frequency parameter plays more important role as compared to acoustic frequency parameter, as it affects the classification accuracy more significantly. One of the limitations of present work is that, we have worked on manually segmented fricative segments. In future, we would like to extend this work on continuous speech.

Acknowledgments. The authors would like to thank Department of Electronics and Information Technology (DeitY), New Delhi (India) and authorities of DA-IICT, Gandhinagar for their support to carry out this research work. We would also like to thank Mr. Maulik Madhavi for his support in the preparation of the paper.

References

1. Quatieri, T.F.: Discrete-time Speech Signal Processing: Principles and Practice. Prentice Hall Press, Upper Saddle River (2004)
2. Web Source, `http://www.langsci.ucl.ac.uk/ipa/IPA_chart_C2005.pdf` (last accessed on 30th April, 2013)
3. Garofolo, J.S.: Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database. National Institute of Standards and Technology (NIST), Gaithersburgh, MD (1988)
4. Scanlon, P., Ellis, D., Reilly, R.: Using Broad Phonetic Group Experts for Improved Speech Recognition. IEEE Trans. on Audio, Speech and Language Proc. 15, 803–812 (2007)
5. Ali, A.M.A., Spiegel, J.V., Mueller, P.: Acoustic-phonetic features for automatic classification of fricatives. J. Accoust. Soc. of America 109(5), 2217–2235 (2001)
6. Ali, A.M.A., Spiegel, J.V., Muller, P.: An acoustic-phonetic feature-based system for the automatic recognition of fricative consonents. In: IEEE Proc. on Int. Conf. on Acoustics, Speech and Signal Processing, vol. 2, pp. 961–964 (1998)
7. Seneff, S.: A Joint Synchrony/ Mean Rate Model of Auditory Speech Processing. J. Phonetics 16, 55–76 (1988)
8. Atlas, L., Shamma, A.S.: Joint acoustic and modulation frequency. EURASIP J. on Applied signal Proccessing 7, 668–675 (2003)
9. Greenberg, S., Kingsbury, B.: The modulation spectrogram: In pursuit of an invariant representation of speech. In: IEEE Proc. on Int. Conf. on Acoust., Speech, Signal Process., Munich, Germany, vol. 3, pp. 1647–1650 (1997)
10. Markaki, M., Stylianou, Y.: Voice pathology detection and discrimination based on modulation spectral features. IEEE Trans. on Audio, Speech, and Language Proc. 19(7), 1938–1948 (2011)
11. Lathauwer, L.D., Moor, B.D., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. 21(4), 1253–1278 (2000)
12. Modulation Toolbox, `http://www.ee.washington.edu/research/isdl/projects/modulationtoolbox` (last accessed on 30th April 2013)