Perspectives in Science (2016) 8, 28—31

ScienceDirect

journal homepage: www.elsevier.com/pisc

Hindi vowel classification using QCN-MFCC features☆

Shipra Mishra*, Anirban Bhowmick, Mahesh Chandra Shrotriya

Electronics and Communication Engineering Department, BIT Mesra, Ranchi 835215, India

Received 18 January 2016; accepted 25 January 2016
Available online 2 March 2016

KEYWORDS
Speech recognition;
Lombard effect;
QCN;
Hidden Markov Model (HMM);
Mel Frequency Cepstral Coefficients (MFCC)

Summary In presence of environmental noise, speakers tend to emphasize their vocal effort to improve the audibility of voice. This involuntary adjustment is known as Lombard effect (LE). Due to LE the signal to noise ratio of speech increases, but at the same time the loudness, pitch and duration of phonemes changes. Hence, accuracy of automatic speech recognition systems degrades. In this paper, the effect of unsupervised equalization of Lombard effect is investigated for Hindi vowel classification task using Hindi database designed at TIFR Mumbai, India. Proposed Quantile-based Dynamic Cepstral Normalization MFCC (QCN-MFCC) along with baseline MFCC features have been used for vowel classification. Hidden Markov Model (HMM) is used as classifier. It is observed that QCN-MFCC features have given a maximum improvement of 5.97% and 5% over MFCC features for context-dependent and context-independent cases respectively. It is also observed that QCN-MFCC features have given improvement of 13% and 11.5% over MFCC features for context-dependent and context-independent classification of mid vowels.

Introduction

The rapid increase in speech based applications in mobile platforms such as tablets, smart phones, etc. has catalyzed the research in the field of ''speech as a medium of human-machine communication'' in last few decades. A no. of ASR systems with good accuracy in clean laboratory environment has been developed till date. But the recognition accuracy of these systems degrades severely in real time situations due to several factors such as environmental noise, variations in speaking styles due to age, gender, dialect, emotional state, etc. Here we have concentrated on how Lombard effect affects the performance of ASR. LE may significantly degrade the performance of ASR systems as it affects a no. of speech production parameters (Bořil and Hansen, 2010; Bořil, 2008; Callejas and López-Cózar, 2008; Bořil et al., 2010). Here we have focused on the classification of Hindi vowels in three classes: front vowel,

☆ This article belongs to the special issue on Engineering and Material Sciences.

* Corresponding author. Tel.: +91 9939874636.
E-mail addresses: mishra.shipra.bit@gmail.com (S. Mishra), anirban.bhowmick@outlook.com (A. Bhowmick), shrotriya@bitmesra.ac.in (M.C. Shrotriya).

http://dx.doi.org/10.1016/j.pisc.2016.01.010

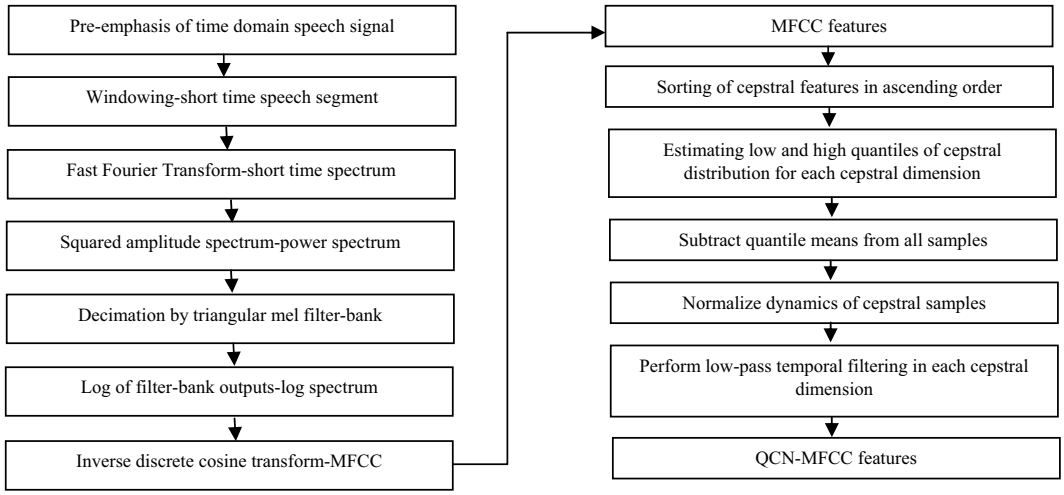

Figure 1 Block diagram of MFCC and QCN-MFCC feature extraction technique.

mid vowel, back vowel. HMM is found to be very effective for phoneme classification tasks (Biswas et al., 2014a,b; Davis and Mermelstein, 1980). Hence we have used HMM as classifier. MFCC has been used as baseline feature extraction technique as it is based on speech perception mechanism in humans (Davis and Mermelstein, 1980). A new feature set QCN-MFCC is prepared by applying equalization of LE on MFCC features (Bořil and Hansen, 2010; Bořil et al., 2006). A Hindi speech database (Samudravijaya et al., 2002) has been used for feature extraction. The rest of the paper is organized as follows. Second section gives the details of the Hindi speech database. Third section gives an overview of feature extraction techniques. Fourth section briefly discusses the acoustic-phonetic features of Hindi language. The comparative recognition efficiency of the two feature sets for Hindi vowel classification is presented and discussed in the last two sections.

Hindi speech database

For phoneme extraction a Hindi Speech Database (Samudravijaya et al., 2002) is used. The database has 100 speakers, each speaker has spoken 10 phonetically rich sentences. Two sentences are common to all the speakers. These sentences are designed at TIFR, Mumbai, India and recorded at CEERI, New Delhi, India. The sampling frequency has been kept 16 kHz, and the speech data is stored in the 16-bit PCM-encoded waveform format in mono-mode. The speech data is recorded using two microphones: one good quality close talking directional microphone and another desk-mounted omni-directional microphone kept at a distance of 1 m from the speaker. In this database manually segmented phoneme boundaries are provided from the spoken sentences. From this database, a total of 50 speakers are chosen out of which 33 speakers are male and 17 speakers are females.

Feature extraction

In this work the performance of QCN-MFCC features (Bořil and Hansen, 2010; Bořil, 2008; Bořil et al., 2006) over baseline MFCC features (Davis and Mermelstein, 1980) is evaluated for Hindi Vowel Classification task. The flow chart representation of the steps involved in MFCC and QCN-MFCC feature extraction technique is given by Fig. 1.

The zeroth cepstral coefficient of MFCC contains the energy of speech signal. The first and second cepstral coefficients display the spectral slope of glottal waveforms (Boril and Hansen, 2011). The formant configurations are displayed by higher cepstral coefficients. The variation in speech due to LE directly affects the cepstra of short time speech segments. Thus it is a logical conclusion that the robustness of any cepstrum based feature extraction technique such as MFCC can be improved by using techniques for equalization of Lombard effect. In this work Quantile-based Dynamic Cepstral normalization technique is applied on MFCC features resulting in a new feature set, QCN-MFCC. The experimental results verified that the QCN-MFCC features gave much better results for Hindi vowel classification task compared to MFCC features for both context dependent and context independent cases. The concept of QCN has evolved from the concept of cepstral mean normalization (CMN) and, cepstral variance normalization (CVN) and hence takes care of the limitations of these two techniques. QCN exploits the cepstral histogram quantiles for determining the dynamic range of the cepstral samples. The quantile means are subtracted from all samples and then the dynamic range is normalized to unity. Then low-pass temporal filtering is applied to improve robustness of features.

Acoustic-phonetic feature of Hindi

In this work QCN-MFCC features are used for classification of Hindi vowels with HMM classifier. The motivation lies in the fact that Hindi is spoken by around 310 million people across the world. And in spite of nearly two decades of research in the field of Hindi speech recognition, a Hindi ASR system with good accuracy in real time situations is yet to be developed. Linguistically the acoustic-phonetic features of Hindi are very different from the European languages. The Hindi alphabet consists of 10 vowels, 4 semivowels, 4 fricatives and 25 stop consonants (Samudravijaya et al., 2002). Here,

Table 1 Hindi language acoustic classes and their phoneme members.

Vowels	Front	इ	ई	ए	ऐ
	Middle	अ	आ		
	Back	उ	ऊ	ओ	औ

we have emphasized Hindi vowels. There are 10 vowels in Hindi Alphabet. The vowels are of three types; front vowel, mid vowel and, back vowel. The vowel classification of Hindi phonemes as given by Samudravijaya et al. (2002) is given in Table 1.

Experimental setup

A Hindi speech database (Samudravijaya et al., 2002) is used for phoneme classification. From this database a total no. of 50 speakers having 17 female speakers and 33 male speakers is taken. Training of system is done with 40 speakers and testing of system is done with 10 speakers. The efficiency of system has been tested with HMM as classifier. Here two types of features are extracted for vowel classification task, MFCC features and QCN-MFCC features (Fig. 2).

One HMM model is prepared for front vowel, mid vowel and back vowel each having 3 emitting states and 4 Gaussian Mixture Components, with spherical covariance. Accuracy of classification is calculated by following relation:

$$\text{Efficiency} = \frac{((\text{total test samples} - \text{error}) \times 100)}{\text{total samples}}$$

Results

The classification task is carried out using 13 MFCC features and 13 QCN-MFCC features for each Hindi phoneme segmented from the database. The features have been derived with a frame size of 16 ms and overlapping of 10 ms. Hamming window is used for windowing. At first experiment were carried out for context-independent (CI) phoneme classification case, then same experiments was carried out for context-dependent (CD) phoneme classification case. The results obtained are shown in Table 2. As is clear from the results the CD phoneme classification results are better than

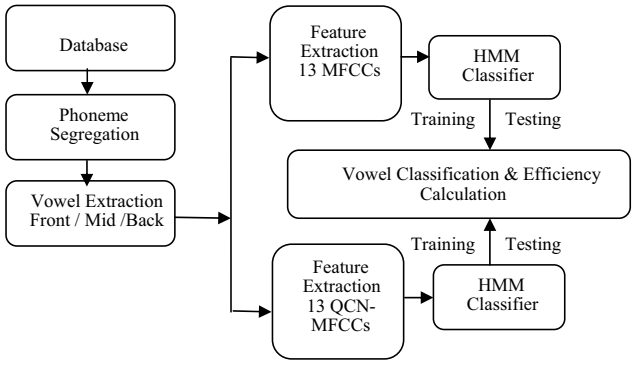

Figure 2 Block diagram of experimental setup.

Table 2 Comparative % recognition efficiency of MFCC and QCN-MFCC features for Hindi vowel classification.

Phoneme	Test set																	
	Clean data				10 dB				5 dB				0 dB					
	CI		CD		CI		CD		CI		CD		CI		CD			
	MFCC	QCN-MFCC	MFCC	QCN-MFCC	MFCC	QCN-MFCC	MFCC	QCN-MFCC	MFCC	QCN-MFCC	MFCC	QCN-MFCC	MFCC	QCN-MFCC	MFCC	QCN-MFCC		
Front	92	93	95	96	67.2	67.98	69.82	72.1	56.91	57.83	58.95	61.85	43.86	44.2	46	47.03		
Mid	84.5	96	85.5	98.5	64.18	72.2	65.12	74.81	54.05	61.98	55.92	65.1	41.9	48	44.87	49.8		
Back	86.5	89	88.5	92.5	65.88	66.34	66.92	67.25	55.28	56.5	56.45	57.83	43.23	44.81	44.25	46.25		
Avg	87.6	92.7	89.7	95.7	65.75	68.84	67.28	71.38	55.41	58.77	57.11	61.59	42.99	45.67	45.04	47.69		

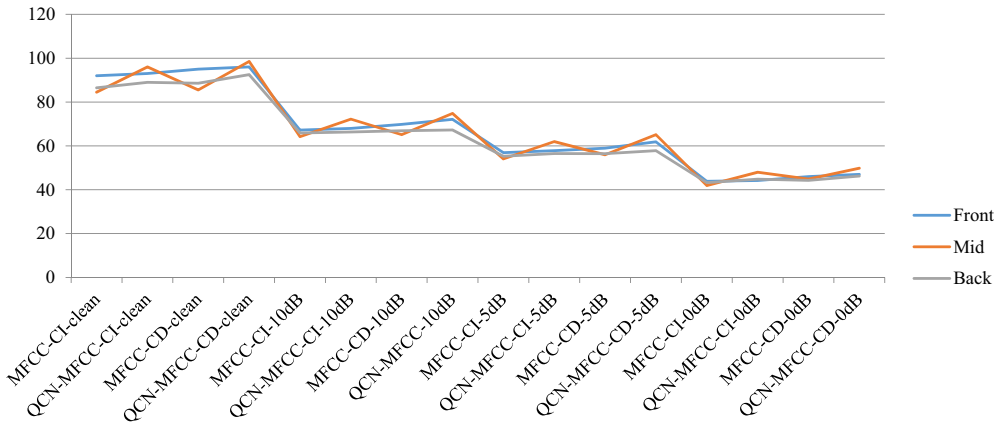

Figure 3 Comparative % recognition efficiency of MFCC and QCN-MFCC features for Hindi vowel classification.

the CI phoneme classification. This improvement in classification accuracy can be due to the fact that Hindi phones are highly affected by their neighbouring phonemes. It is also clear from the results that QCN-MFCC features gave better recognition efficiency in all the cases. This improvement can be explained by the fact that QCN is an unsupervised normalization technique which compensates for the mismatch between acoustic models and test speech samples.

Here an HMM classifier is used for classifying Hindi vowels in three fundamental categories; front vowel, mid vowel and back vowel. Two feature extraction techniques are used for this purpose; MFCC and QCN-MFCC features. 13 cepstral coefficients are taken for each feature extraction technique. It is observed that QCN-MFCC features give better recognition performance for both context-independent and context-dependent cases. This may be because of the fact that QCN technique compensates for the mismatch between acoustic models and speech signal. It is also observed that CD results were better in all cases than CI results. This may be due to the inter-dependency of neighbouring phonemes (Fig. 3).

References

Biswas, A., Sahu, P., Chandra, M., 2014a. Admissible wavelet packet features based on human inner ear frequency response for Hindi consonant recognition. Comput. Electr. Eng. 40 (4), 1111—1122.

Biswas, A., Sahu, P., Bhowmick, A., Chandra, M., 2014b. Feature extraction technique using ERB like wavelet sub-band periodic and aperiodic decomposition for TIMIT phoneme recognition. Int. J. Speech Technol. 17, 389—399.

Bořil, H., (Ph.D. thesis) 2008. Robust Speech Recognition: Analysis and Equalization of Lombard Effect in Czech Corpora. Czech Technical University in Prague, Czech Republic http://www.utdallas.edu/hynek.

Boril, H., Hansen, J.H.L., 2011. UT-scope: towards LVCSR under Lombard effect induced by varying types and levels of noisy background. In: IEEE ICASSP'11, Prague, Czech Republic, May 2011, pp. 4472—4475.

Bořil, H., Hansen, J.H.L., 2010. Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments. IEEE Trans. Audio Speech Lang. Process. 18 (August (6)), 1379—1393.

Bořil, H., Fousek, P., Pollák, P., 2006. Data-driven design of front-end filter bank for Lombard speech recognition. In: Proc. ICSLP'06, Pittsburgh, PA, pp. 381—384.

Bořil, H., Sadjadi, O., Kleinschmidt, T., Hansen, J.H.L., 2010. Analysis and detection of cognitive load and frustration in drivers' speech. In: Interspeech'10, Makuhari, Japan, September 2010, pp. 502—505.

Callejas, Z., López-Cózar, R., 2008. Influence of contextual information in emotion annotation for spoken dialogue systems. Speech Commun. 50 (5), 416—433.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28, 357—366.

Samudravijaya, K., Rao, P.V.S., Agrawal, S.S., 2002. Hindi speech database. In: Proc. Int. Conf. on Spoken Language Processing (ICSLP00), Beijing, China, October 2002, pp. 456—459.