Recognition of Subsampled Speech using a Modified Mel Filter Bank

Kiran Kumar Bhuvanagiri, Sunil Kumar Kopparapu

TCS Innovation Labs - Mumbai,
Tata Consultancy Services, Pokhran Road 2, Thane (West),
Maharastra 400 601. INDIA.
{kirankumar.bhuvanagiri, sunilkumar.kopparapu}@tcs.com

Abstract. Several speech recognition applications use Mel Frequency Cepstral Coefficients (MFCCs) . In general, these features are used to model speech in the form of HMM. However, features depend on the sampling frequency of the speech and subsequently features extracted at certain rate can not be used to recognize speech sampled at a different sampling frequency [5]. In this paper, we first propose a modified Mel filter bank so that the features extracted at different sampling frequencies are correlated. We show experimentally that the models built with speech sampled at one frequency can be used to recognize subsampled speech with high accuracies.

Keywords: MFCC, speech recognition, subsampled speech recognition

1 Introduction

Mel Frequency Cepstral Coefficients (MFCC) are commonly used features in speech signal processing. They have been in use for a long time [3] and have proved to be one of the most successful features in speech recognition tasks [8]. For a typical speech recognition process (see Fig. 1), acoustic models are built using speech recorded at some sampling frequency during the training phase (boxed blue -.-. in Fig. 1). In the testing (boxed red - - - in Fig. 1) or the recognition phase, these acoustic models are used along with a pronunciation lexicon and a language model to recognize speech at the *same* sampling frequency. If the speech to be recognized is at a sampling frequency other than the sampling frequency of the speech used during the training phase then one of the two things needs to be done (a) *retrain* the acoustic models with speech samples of the desired sampling frequency or (b) change the sampling rate of the speech to be recognized (test speech) to match the sampling frequency of the speech used for training. In this paper, we address the problem of using models built for a certain sampling frequency to enable recognition of speech at a different sampling frequency. We particularly concentrate on Mel-frequency cepstral coefficient (MFCC) as features [9], [4] because of their frequent use in speech signal processing. Kopparapu et al [5] proposed six filter bank constructs to enable

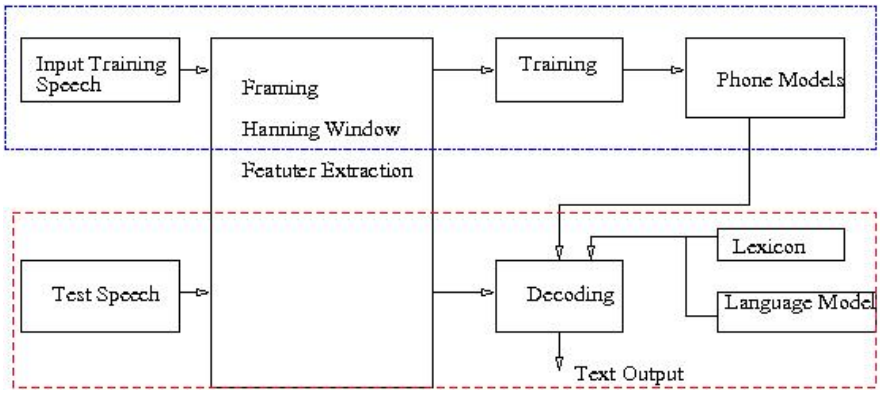

Fig. 1. Speech Recognition - showing train and the test stages.

calculation of MFCC's of a subsampled speech. Pearson correlation coefficient was used to compare the MFCC of subsampled speech and the MFCC of original speech.

In this paper, we construct a Mel filter bank that is able to extract MFCCs of the subsampled speech which are significantly correlated to the MFCC's of original speech compared to the Mel filter banks discussed in [5]. This is experimentally verified in two ways (a) through the Pearson correlation coefficient and (b) through speech recognition experiments on AN4 speech database [1] using open source ASR engine [2]. Experimental results show that the recognition accuracy on subsampled speech using models developed using original speech is as good as the recognition accuracy on original speech and as expected degrades with excessive subsampling.

One of the prime applications of this work is to enable use of acoustic models created for desktop speech (usually 16 kHz) with telephone speech (usually 8 kHz) especially when there is access to only the acoustics models and not to the speech corpus specifically as in SPHINX. The rest of the paper is organized as follows. In Section 2, largely based on our previous work [5], procedure to compute MFCC features and the relationship between the MFCC parameters of the original and subsampled speech is discussed. In Section 2.1, new filter bank is proposed. Section 3 gives the details of the experiments conducted to substantiate advantage of proposed modified filter bank and we conclude in Section 4.

2 Computing MFCC of Subsampled Speech

As shown in [5], let $x[n]$ be a speech signal with a sampling frequency f_s and be divided into P frames each of length N samples with an overlap of $N/2$ samples, say, $\{\boldsymbol{x}_1, \boldsymbol{x}_2 \cdots \boldsymbol{x}_p \cdots \boldsymbol{x}_P\}$, where $\boldsymbol{x}_p$ denotes the p^{th} frame of the speech signal

$x[n]$ and is $\boldsymbol{x}_p = \left\{x\left[p*\left(\frac{N}{2}-1\right)+i\right]\right\}_{i=0}^{N-1}$. Computing MFCC of the p^{th} frame involves,

1. Multiply $\boldsymbol{x}_p$ with a hamming window $w[n] = 0.54 - 0.46\cos\left(\frac{n\pi}{N}\right)$,
2. Compute discrete Fourier transform (DFT) [7]. Note that k corresponds to the frequency $l_f(k) = kf_s/N$.

$$X_p(k) = \sum_{n=0}^{N-1} x_p[n]w[n]\exp^{-j\frac{2\pi kn}{N}} \quad \text{for} \quad k = 0, 1, \cdots, N-1$$

3. Extract the magnitude spectrum $|X_p(k)|$
4. Construct a Mel filter bank $M(m,k)$, typically, a series of overlapping triangular filters defined by their center frequencies $l_{fc}(m)$. The parameters that define a Mel filter bank are (a) number of Mel filters, F, (b) minimum frequency, l_{fmin} and (c) maximum frequency, l_{fmax}. So, $m = 1, 2, \cdots, F$ in $M(m,k)$.
5. Segment the magnitude spectrum $|X_p(k)|$ into F critical bands by means of a Mel filter bank.
6. The logarithm of the filter bank outputs is the Mel filter bank output

$$L_p(m) = ln\left\{\sum_{k=0}^{N-1} M(m,k)|X_p(k)|\right\} \tag{1}$$

   where $m = 1, 2, \cdots, F$ and $p = 1, 2, \cdots, P$.
7. Compute DCT of $L_p(m)$ to get the MFCC parameters.

$$\Phi_p^r\{x[n]\} = \sum_{m=1}^{F} L_p(m)\cos\left\{\frac{r(2m-1)\pi}{2F}\right\} \tag{2}$$

   where $r = 1, 2, \cdots, F$ and $\Phi_p^r\{x[n]\}$ represents the r^{th} MFCC of the p^{th} frame of the speech signal $x[n]$.

The sampling of the speech signal in time effects the computation of MFCC parameters. Let $y[s]$ denote the sampled speech signal such that $y[s] = x[\alpha n]$ where $\alpha = \frac{u}{v}$ and u and v are integers. Note that $\alpha > 1$ denotes downsampling while $\alpha < 1$ denotes upsampling and for the purposes of analysis we will assume that α is an integer. Let $y_p[s] = x_p[\alpha n]$ denote the p^{th} frame of the time scaled speech where $s = 0, 1, \cdots, S-1$, S being the number of samples in the time scaled speech frame given by $S = N/\alpha$. DFT of the windowed $y_p[n]$ is calculated from the DFT of $x_p[n]$. Using the scaling property of DFT, we have, $Y_p(k') = \frac{1}{\alpha}\sum_{l=0}^{\alpha-1} X_p(k'+lS)$ where $k' = 1, 2, \cdots, S$. The MFCC of the subsampled speech is given by

$$\Phi_p^r\{y[n]\} = \sum_{m=1}^{F} L'_p(m)cos\left\{\frac{r(2m-1)\pi}{2F}\right\} \tag{3}$$

where $r = 1, 2, \cdots, F$ and

$$L'_p(m) = ln \left\{ \sum_{k'=0}^{S-1} M'(m,k') \left| \frac{1}{\alpha} \sum_{l=0}^{\alpha-1} X_p(k' + lS) \right| \right\} \quad (4)$$

Note that L'_p and M' are the log Mel spectrum and the Mel filter bank of the subsampled speech. Note that a good choice of $M'(m,k')$ is the one which gives (a) the best Pearson correlation with the MFCC ($M(m,k)$ of the original speech and (b) best speech recognition accuracies when trained using the original speech and decoded using the subsampled speech. Kopparapu et al [5] chose different constructs of $M'(m,k')$.

2.1 Proposed Filter Bank

We propose a Mel filter bank $M_{new}(m,k')$ for subsampled speech as

$$M_{new}(m,k') = \begin{cases} M(m,\alpha k') & \text{for } \; l_f(k') \leq (\frac{1}{\alpha}\frac{f_s}{2}) \\ 0 & \text{for } \; l_f(k') > (\frac{1}{\alpha}\frac{f_s}{2}) \end{cases}$$

where k' ranges from 1 to N/α. Notice that the modified filter bank is the subsampled version of original filter bank with bands above $\frac{1}{\alpha}\frac{f_s}{2}$ set to 0. Clearly, the number of Mel filter bands are less than the original number of Mel filter bands. Let ξ be the number of filter banks whose f_c is below $f_s/2\alpha$. Subsequently, $L_p(m)$, the Mel filter bank output, is 0 for $m > \xi$. In order to retain the total number of filter bank outputs as the original speech we construct

$$L_p(m) = (0.9)^{m-\xi-1} L_p(\xi - 1) \;\text{ for }\; \xi < m \leq F \quad (5)$$

Equation (5) is based on the observation that Mel filter outputs for $m > \xi$ seems to decay exponentially.

3 Experimental Results

We conducted experiments on AN4 [1] audio database. It consists of 948 train and 130 test audio files containing a total of 773 spoken words or phrases. The recognition results are based on these 773 words and phrases. All the speech files in the AN4 database are sampled at 16 kHz. The Mel filter bank has $F = 30$ bands with $l_{fmin} = 130$ Hz and $l_{fmax} = 7300$ Hz and the frame size is set to 32 ms. The MFCC parameters are computed for the 16 kHz speech signal $x[n]$, and also the subsampled speech signal $x[\alpha n]$. The MFCC parameters of $y[s]$ are calculated using the proposed Mel filter bank (5) while the MFCC of $x[n]$ was calculated using (2).

We conducted two types of experiments to evaluate the performance of the proposed Mel filter bank construction on subsampled speech. In the first set of experiments, we used the Pearson correlation coefficient (r) to compare the MFCC of the subsampled speech with the MFCC of the original speech along the

lines of [5]. In the second set of experiments we used speech recognition accuracies to evaluate the appropriateness of the use of Mel filter bank for subsampled speech. We compared our Mel filter bank with the best Mel filter bank (Type C) proposed in [5].

3.1 Comparison using Pearson correlation coefficient

We computed framewise r for MFCCs of subsampled speech and the original speech. The mean and variances of r over all the frames are shown in Table 1. Clearly the Mel filter bank construction proposed in this paper performs better than the best method suggested in [5] for all values of α. For $\alpha = 16/4 = 4$ the mean-variance pair for the proposed Mel filter bank is $(0.85609, 0.04176)$ compared to best in [5] $(0.67837, 0.14535)$.

Table 1. Pearson correlation coefficient (r), between the MFCCs of original and subsampled speech

Sampling factor	**Proposed**		**best in [5]**	
	mean	variance	mean	variance
$\alpha = 16/4$	0.85609	0.04176	0.67837	0.14535
$\alpha = 16/5$	0.90588	0.02338	0.70064	0.1280
$\alpha = 16/6$	0.9284	0.01198	0.7201	0.1182
$\alpha = 16/7$	0.94368	0.00633	0.7321	0.1010
$\alpha = 16/8$	0.96188	0.00005	0.7465	0.0846
$\alpha = 16/10$	0.98591	0.00037	0.8030	0.0448
$\alpha = 16/12$	0.989	0.00025	0.8731	0.0188
$\alpha = 16/14$	0.99451	0.00006	0.9503	0.0029
$\alpha = 16/16$	1	0	1	0

3.2 Speech recognition Experiments

We used the 948 training speech samples of AN4 database to build acoustic models using SphinxTrain. Training is done using MFCCs calculated on the 16 kHz (original) speech files. Recognition results are based on the 130 test speech samples In Case A we used 30 MFCC's while in Case B we used 13 MFCC but concatenated them with 13 velocity and 13 acceleration coefficients to form a 39 dimensional feature vector. Recognition accuracies are shown in Table 2 on the 773 words in the 130 test speech files. It can be observed that the word recognition accuracies using the proposed Mel filter bank on subsampled speech is better that what has been proposed in [5] for all values of α and for both the Case A and the Case B. We also observe from Fig. 2 that proposed method is more robust, while accuracies rapidly fall in case of best method in [5] it gradually decreases in our case.

Table 2. Recognition accuracies (percentage)

Sampling factor	Case A (30 MFCCs)		Case B (39 features)	
	proposed	Best in [5]	proposed	Best in [5]
$\alpha = 16/4$	9.83	2.07	30.36	3.36
$\alpha = 16/5$	20.18	1.68	58.86	2.85
$\alpha = 16/6$	27.30	2.07	68.95	3.62
$\alpha = 16/7$	31.44	2.33	73.22	5.30
$\alpha = 16/8$	37	3.88	77.23	11.77
$\alpha = 16/10$	40.36	7.12	80.50	34.15
$\alpha = 16/12$	41.01	16.19	81.11	65.85
$\alpha = 16/14$	42.56	34.80	82.54	77.10
$\alpha = 16/16$	43.21	43.21	81.11	81.11

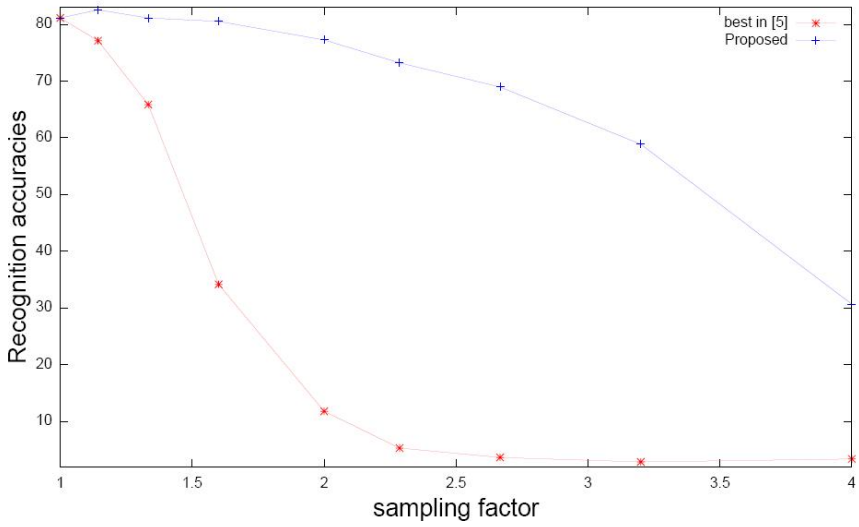

Fig. 2. Comparing ASR accuracies of both methods for different values of sampling factors (α).

The better performance of the proposed Mel filter bank in terms of recognition accuracies can be explained by looking at a sample filter bank output shown in Fig. 3. Filter bank output of the proposed Mel filter bank construct (red line '+') closely follow that of original speech Mel filter bank output (blue line 'x'), while even the best reported filter bank in [5] (shown in black line 'o') shows a shift in the filter bank outputs.

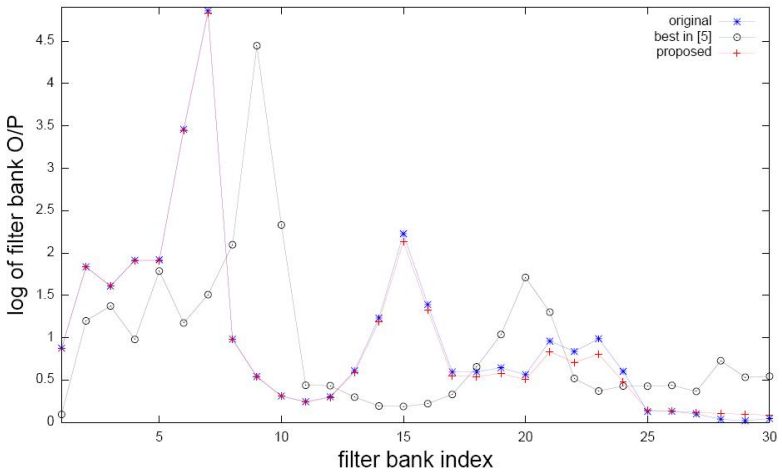

Fig. 3. Sample log Filter bank outputs of original speech, and subsampled speech using the proposed Mel filter bank and the best Mel filter bank in [5]

4 Conclusion

The importance of this Mel filter bank design to extract MFCC of subsampled speech is apparent when there are available trained models for speech of one sampling frequency and the recognition has to be performed on subsampled speech without explicit creation of acoustic models for the subsampled speech. As a particular example, the work reported here can be used to recognize subsampled speech using acoustic (HMM or GMM) models generated using Desktop speech (usually 16 kHz). We proposed a modified Mel filter bank which enables extraction of MFCC from subsampled speech which correlated very well with the MFCC of the original sampled speech. We experimentally showed that the use of the modified Mel filter bank construct in MFCC computation of subsampled speech outperforms the Mel filter banks developed in [5]. This was demonstrated at two levels, namely, in terms of a correlation measure with the MFCC of the original speech and also through word recognition accuracies. Speech recognition accuracies for larger values of α can be improved by better approximating the

missing Mel filter outputs using bandwidth expansion [6] techniques, which we would be addressing in our future work.

References

1. CMU: AN4 database, http://www.speech.cs.cmu.edu/databases/an4/
2. CMU: Sphinx, http://www.speech.cs.cmu.edu/
3. Davis, S.B., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Processing 28(4), 357–366 (1980)
4. Jun, Z., Kwong, S., Gang, W., Hong, Q.: Using Mel-frequency cepstral coefficients in missing data technique. EURASIP Journal on Applied Signal Processing 2004, no. 3, 340–346 (2004)
5. Kopparapu, S., Laxminarayana, M.: Choice of mel filter bank in computing mfcc of a resampled speech. In: Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on. pp. 121 –124 (May 2010)
6. kornagel, U.: Techniques for artificial bandwidth extension of telephone speech. Signal processing 86, Issue 6 (June 2006)
7. Oppenheim, Schafer: Discrete Time Signal Processing. Prentice-Hall (1989)
8. Quatieri, T.F.: Discrete-time speech signal processing: Principles and practice. Pearson Education II, 686, 713 (1989)
9. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing 3, No. 1 (January 1995)