EVALUATION OF SPEAKER RECOGNITION FEATURE-SETS USING THE SVM CLASSIFIER

Daniel J. Mashao and John Greene

STAR Research Group, Department of Electrical Engineering,
UCT, Rondebosch, 7701, South Africa,
www.star.za.net
daniel@eng.uct.ac.za & jrgreene@eng.uct.ac.za

ABSTRACT

Feature-sets play an important role in the performance of speaker recognition systems. Design of optimal feature-sets is still an ongoing research effort. One characteristic of feature-sets that is known to have an impact on the performance of a speech technology system is the compression that is applied during the computation of the feature-sets in the spectral domain. Most systems use the mel-scale compression via the mel-spaced triangular filters. The mel-scale was designed from studying the response of the basilar membrane in the inner ear to external frequency stimuli. In this paper we investigated the impact of different levels of feature-sets spectral compression on the performance of the support vector machine (SVM) classifier for speaker identification task. We found that spectral compression does impact performance, and that the best performance is obtained above the mel-scale compression. This confirms results in literature that shows better feature-set performance is above mel-scale compression for male speakers, although the literature results were done in speech recognition and not speaker recognition. The NTIMIT database New England region was used for the experiments and the gender spread favoured males.

Keywords: speaker recognition, SVM, feature-sets, mel-scale

1. INTRODUCTION

Speech technology systems uses a speaker's voice to operate. In speaker recognition systems the aim is to extract features from the voice that are independent of the words (text) that the speaker is saying. This differs from speech recognition task where the aim is to extract features from the speech signal that are ideally independent of a specific speaker. In both cases the aim is to extract features that are relevant to the task, while in speaker recognition the words form part of the noise in the system, for speech recognition the individual's characteristics are part of the noise. The most common method used for feature extraction in speech technology research is called the mel-frequency cepstral coefficients (MFCC). This will be discussed further in section 3 of this document. The mel refers to the compression that is applied on the feature-sets and the cepstrum refers to the fact that the features are extracted in the cepstral domain. The mel is based on studies of the response of the nerve cells along the basilar membrane in the inner ear. Even though the speaker and speech recognition tasks have different goals, in many state of the art systems the same MFCC feature-sets are used for both. This indicates that the features retains the individuals characteristics and the semantics of the spoken text.

There has been some research work on evaluating speaker recognition systems using the support vector machine (SVM) classifier. Most of the work is in speech recognition and usually the classifier is used in conjunction with the generative classifier such as the Gaussian mixture models [3, 2, 1, 5]. The SVM approach is generally more expensive to compute than the GMM models. As we have shown in [4] the SVM does outperform the GMM when there is limited data. With more data the GMM approach is better. This can be easily explained, since with the GMM the more data is available the better are the estimates of means and covariances of the speakers features, whereas in the SVM case more data may not change the support vectors.

The impact of using different spectral compression levels in the feature-set has not been thoroughly investigated. This is mainly because of the risk of designing a spectral compression feature system that fits a particular dataset (over-fitting problem), gender sensitivity and most importantly the expensive computational resources that are required for such experimentation.

In this paper the impact of spectral compression on the performance of the Support Vector Machine (SVM) classifier is investigated. The SVM has emerged in recent years as a classification technique resting on a firm theoretical basis of Statistical Learning Theory, and also exhibiting excellent empirical performance over a wide range of classification problems. How it performs in a speaker identification task when the spectral compression changes is of interest.

2. SUPPORT VECTOR MACHINES

SVM classifiers operate in a way which at first may seem counter-intuitive. Ultimately, classification involves dimensionality reduction, but the SVM first performs an implicit nonlinear projection of the data into a feature space of very high, or even infinite, dimensionality, where it finds a linear or hyperplane decision boundary. Over-fitting is controlled (and good generalisation guaranteed) through the principle of structural risk minimisation, as elaborated by Vapnik and others[10]. The empirical risk of misclassification is minimised by maximising the margin between the data points and the decision boundary. In practice this criterion is softened to the minimisation of a cost factor involving both the complexity of the classifier and the degree to which marginal points are misclassified, and the tradeoff between these factors is managed through a parameter (usually designated C) which is tuned through cross-validation procedures.

The nonlinear projection of the data is performed by means of a 'kernel' function. The commonest kernel (and the one used in this work) is the Gaussian (or "radial basis function") kernel, defined as $K(x, y) = e^{-|x-y|^2/(2\sigma^2)}$, where σ is a scale parameter. The decision function is a weighted linear sum of these Gaussian functions located at each data point (in this it resembles an extreme case of a Gaussian mixture model). However the support vector algorithm finds a sparse expansion in which most of the weighting coefficients are zero (only those closest to the decision boundary, the "support vectors" are nonzero). The parameter C mentioned above sets an upper bound on the value of any given coefficient, limiting local distortion of the decision boundary by noisy data point or outliers. Careful tuning of σ and C is needed to optimise the predictive accuracy of the classifier.

The data used in these experiments is 32 dimensions. The scale parameter $\sigma^2 = 32000$ and error margin parameter is the default $C = 100$ in the SVMTorch package which is used for the experiments. The scale parameter was chosen after a limited number of experiments. The values chosen for the SVM kernel exceeds or matches the performance of the other popular classification method, the Gaussian mixture models on the same dataset, with limited training data.

3. PARAMETRIC FEATURE SETS

The usual method of calculating the feature-sets is called the mel-frequency cepstrum coefficients (MFCC). The MFCC feature-sets are computed as shown in Figure 1. The input speech signal is sampled appropriately (that is low pass filtered and sampling at over twice the Nyquist rate), then converted into the spectral domain via a discrete Fourier transform. In the spectral domain the log magnitude of the complex signal is obtained and finally an inverse discrete Fourier transform is performed on the samples. The resulting samples are in the quefrency or cepstral domain and are called cepstral coefficients. Cepstral is a play on the words spectral in reverse. If the inverse discrete Fourier transform was done without the log magnitude section then the samples would have reverted back to the time domain, therefore the units used in the cepstral domain are seconds. The advantage of working in the cepstral domain is that the pitch which is specific to the speaker is separated from the vocal tract parameters (configuration) that are influenced by the words that the person is saying. This is because in the time domain the speech signal is a convolution of the impulse function and the vocal tract parameters $s(t) = w(t) * v(t)$, where $s(t)$is the output speech signal and $w(t)$is the triangular impulse train (usually modelled by a 1/3 rise and 2/3 fall signal) and $v(t)$ is the vocal tract configuration. In the frequency domain $S[\omega] = W[\omega]V[\omega]$. By taking the log magnitude the function is simplified as a sum and the two quantities can be easily separated. The pitch tends to have a higher quefrency than the vocal tract parameters that are changing very slowly (muscles and bone movements) as a result a low pass filter is sufficient to separate the signals. In addition to the Fourier transform it is known that the ear is sensitive to frequencies differently, that is a difference between 500Hz and 800Hz is more noticeable than a difference between 3000Hz and 3300Hz. To capture this sensitivity a mel-scale compression is applied on the feature-sets in the spectral domain. The most popular approximation of the mel scale is by O'Shaughnessy[7] which is

$$F_{mel} = 2595 \log(1 + \frac{f_{in}}{700})$$

where F_{mel} is the frequency in mels and f_{in} is the input frequency in hertz. As discussed by Umesh et al[9] there are many other functional approximations of this compression with minor operational differences although using the triangular filters is the most common way of implementation. An example of the filters are as shown in Figure 2. The filters are linear up to 1000Hz and then become logarithmic in keeping with our hearing. The MFCC contains important characteristics such as the separation of a pitch from the vocal tract configuration. The vocal tract configuration contains the phonemes (basic speech units).

The parametric feature sets (PFS) works similar to the MFCC and can simulate the mel compression when certain combination of parameters is used. It is flexible and easier to visualise the impact of the parameters using the PFS than using the traditional mel-scale. The PFS is defined by two parameters α and β related as in the following formula

$$A \cdot \sum_{i=1}^{\alpha} \beta^{i-1} = \frac{N}{2}$$

where N is the size of the analysis window and A is a constant, which represents the first region of the spectrum. The next region of the spectrum is $A\beta$, $A\beta^2$, and so on. The spectral sampling is the same within a region and since the regions grow bigger exponentially when $\beta > 1$ a spectral compression is therefore created. The α term is the number of divisions or regions. The mel-scale compression can be seen as having many regions given that the first region is linear and thereafter the filters are logarithmic.

The PFS can generate mel-scale compression but it is flexible as it allows other spectral compressions to be evaluated. Two papers have shown how spectral compression can affect the performance of the speech recognition systems [6, 8]. Figure 3 shows different compressions that are obtained when different β parameters are used with $\alpha = 4$. Notice that the mel scale compression is closer to $\alpha = 4, \beta = 2.0$. This is true only because the window of 320 samples is used. Approximations depends on the window length.

The algorithm of the PFS is a follows; Sample the input speech signal as in the case of MFCC parameters. In the spectral domain apply the log magnitude and filter the log spectra using a low pass filter. This removes the high frequencies. Sample the log spectra using the α and β parameters. Figure 3 shows the compression that is achieved and Figure 4 shows the samples that are actually obtained for various α and β parameters. It can be seen that when $\beta = 1$ there is a uniform sampling of the spectra and as β increases the lower frequencies are sampled more than the higher frequencies. Finally like in the MFCC the actual coefficients are obtained in the cepstral domain using the discrete cosine transform.

The PFS has been used in conjunction with the Gaussian Mixture models to produce one of the highest performance on the NTIMIT database.

4. EVALUATIONS

In this section the database used for the experiments and the experiments performed are discussed.

4.1. Database

One problem of doing speech technology research is that it is difficult to compare results since there are so many differences in

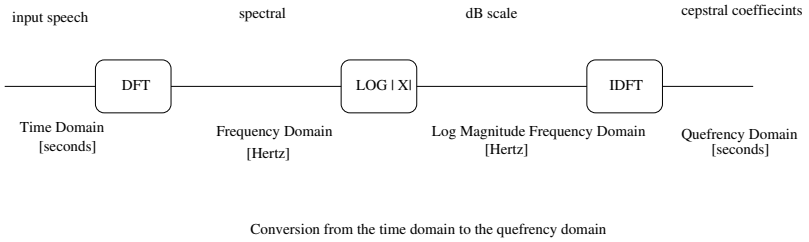

Fig. 1. Conversion from time domain to the quefrency domain

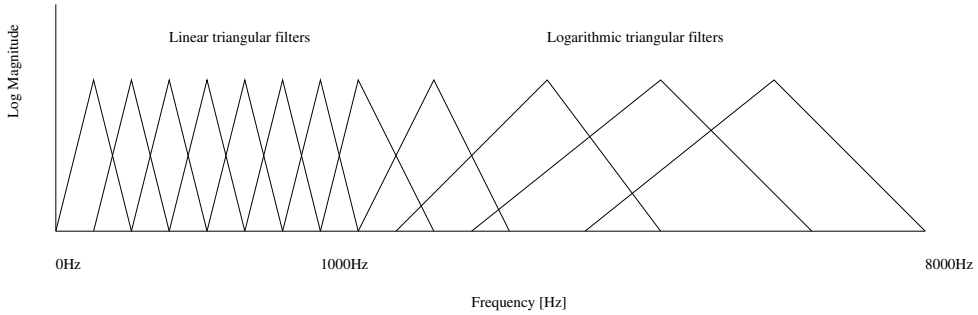

Fig. 2. Filters used to achieve mel-scale compression

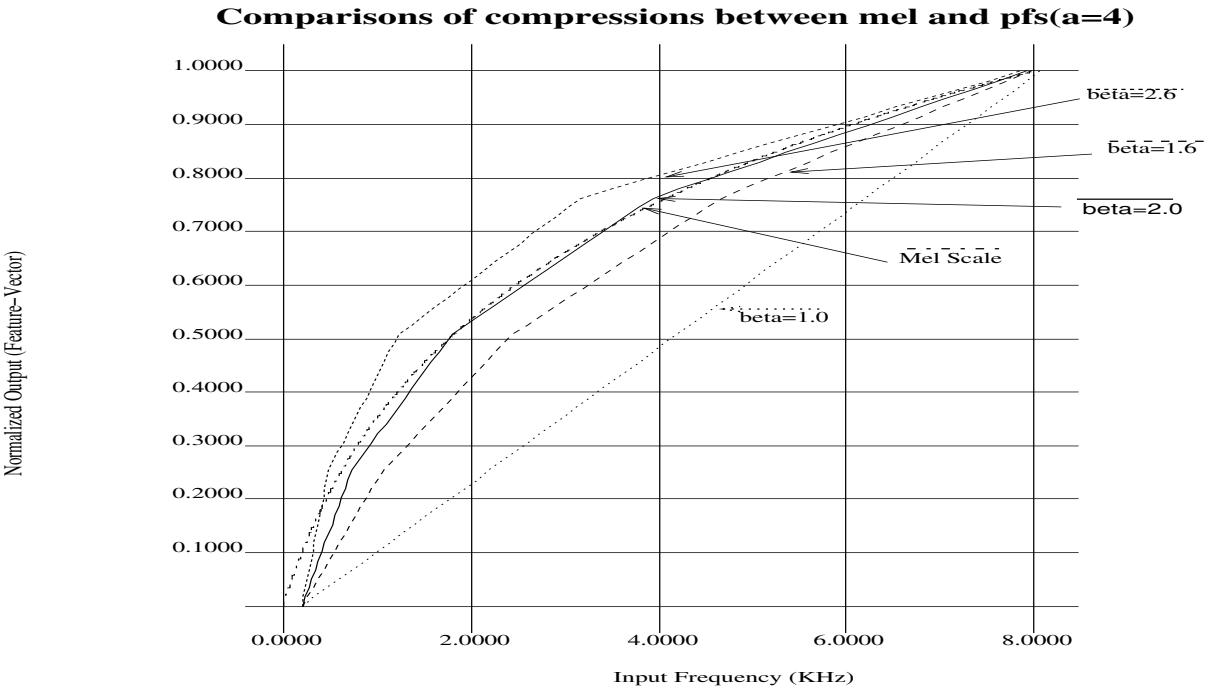

Fig. 3. Comparison between mel-scale compression and PFS with $\alpha = 4$

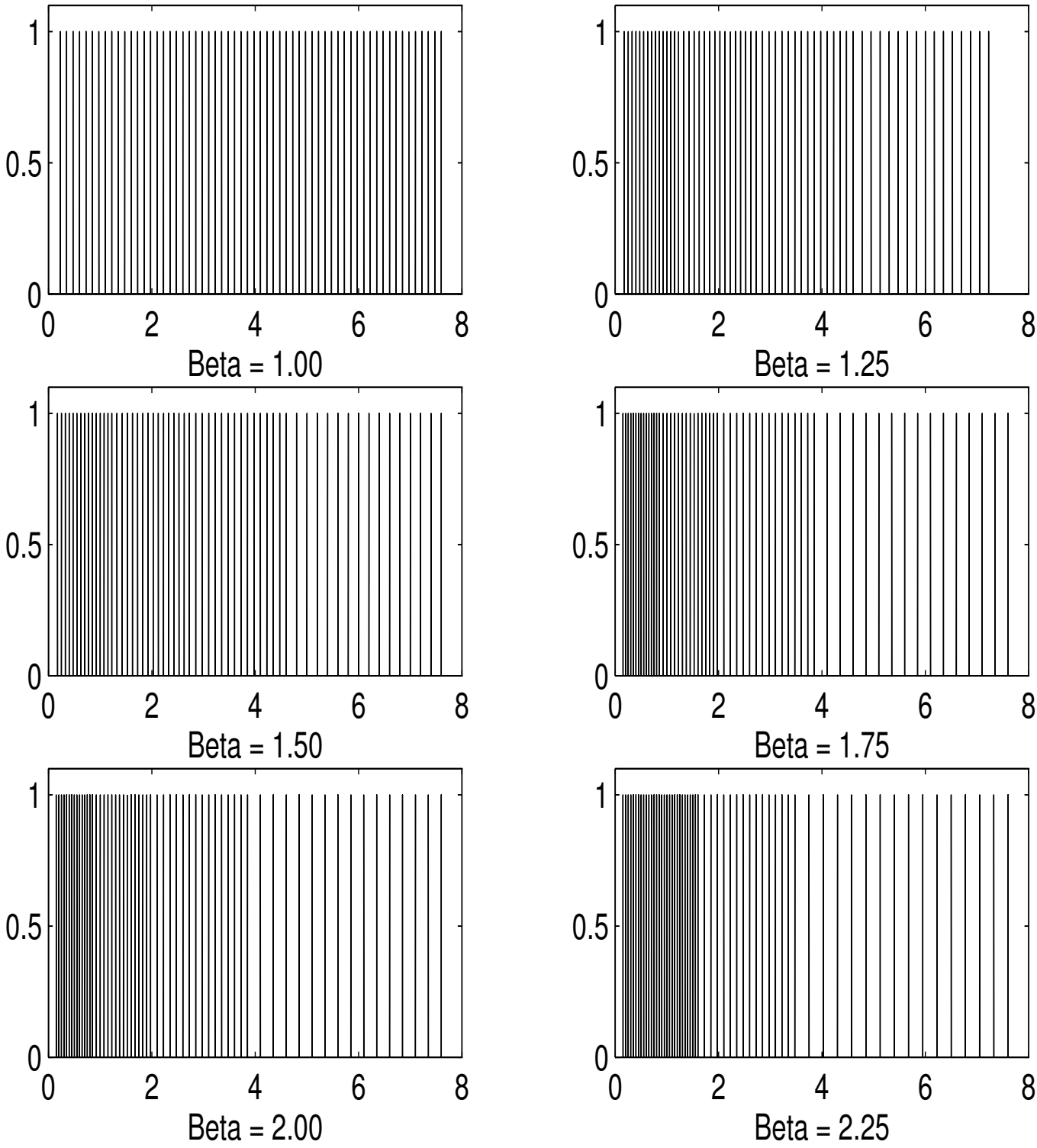

Fig. 4. Sampling of different PFS parameters. X-axis is frequency from 0 to 8K Hz, y-axis is magnitude.

the reported experiments. There are differences in the front-ends, back-ends, preprocessing stages and the database used. The Linguistic Data Consortium (LDC) in the USA is attempting to collect databases that will improve the chances of making valid and meaningful comparisons between different systems. One of the databases we obtained from the LDC is the TIMIT database. TIMIT stands for Texas Instruments (TI) and Massachusetts Institute of Technology (MIT), these are the two organisations that collected the database. The database consists of 630 speakers from all over the USA. The speakers are classed according to the region of the USA where they grew up as this is assumed to be an important factor in determining a person's accent. All the speakers speak American English.

Each speaker spoke ten utterances. All the utterances are different across speakers except utterance 0 and 1, which are common. These common utterances were not used in the results reported in this paper. Utterances are on average three seconds long. Current state of the art speaker identification systems have no problem with the TIMIT database. That is when using 8 utterances to train a speakers' model and just using the other 2 utterances for evaluations, current PFS and MFCC based systems can correctly identify all the 630 speakers in the database. As such the TIMIT database is no longer used for evaluation of speaker identification systems.

The NTIMIT database is based on the TIMIT database, and the N before TIMIT stands for Noisy. The NTIMIT database was passed on the telephone network. It is not a simulation but an actual speech through the telephone network. The best speaker identification performances on the NTIMIT database is around 70% and not 100% as in the case of TIMIT. The NTIMIT database represents a challenge due to unpredictable channel noises. In the reported experiments only the Region 1 train speakers database is used which has 38 speakers. The main hindrance from experimenting with the SVM classifier is that it takes very long time to enrol the speakers (make the models) and to do evaluations.

4.2. Experiments

The problem with SVM is that the computational burden is excessive compared to other competing methods such as the Gaussian mixture models (GMM). The other limitations is extra unknown parameters that are to be specified such as the standard deviation (scale) parameter σ^2 and the margin of error parameter C in the case of the Gaussian kernel as used for this experiments . The experiments took a very long time to complete as such only region dr1 (New England region, train set) of the database was used. To obtain a single value of performance took over 8 hours on a 1.7 GHz Pentium IV machine and it is dependent on the amount of enrolment data.

Two experiments were done. In the first experiment the system was trained with 3 utterances and tested with four utterances. The was repeated 4 times. The results shown are the averages of the performance results, each data point was computed from at least 16 values, except for values from $\beta = 4.8$ which were computed from 12 or less number of values. Table 1 shows which utterances were used in enrolment (training) and which were used in evaluations(testing). There is no particular reason why certain utterances were chosen for testing and training. This was done randomly. The results shown on Figure 5 are the average identification rates of all these tests and short lines show the standard deviation of those results.

Enrol Utterances	Evaluation Utterances
2,3,4	5,6,7,8
4,5,6	2,3,7,8
5,6,9	2,3,7,8
7,2,6	3,4,5,8
3,7,9	2,5,6,8

Table 1. 3 enrol and 4*1 tests

Enrol Utterances	Evaluation Utterances
2,3,7	(4,5), (8,9)
2,3,4	(6,7), (8,9)
4,5,6	(2,3), (8,9)
5,6,7	(3,4), (8,9)
4,5,9	(2,7), (7,8)
3,6,7	(4,5),(8,9)
7,8,9	(3,4), (5,6)

Table 2. 4 enrol and 2*2 tests

For the second experiments the number of training utterances were increased. A similar curve is obtained although it shows higher identification rates as the number of training utterances is four and not three. Also two utterances are used for evaluations at a time instead of one at a time as in the previous experiments. The utterances used for the experiments are shown in Table 2. One may note the eighth and ninth (8,9) utterances are always used for evaluations. This is because of a bias from researchers using the TIMIT databases towards using these last two utterances for evaluations in general[1].

Figure 5 shows the performance when different β values with $\alpha = 4$ fixed. The results show that spectral compression does indeed have an impact on the feature-sets. In the case of no spectral compression($\alpha = 4, \beta = 1.0$) the identification rate of the system was 41.6% (50.4% for the second experiment). As soon as compression is added the performance increased to just under 47%. As compression is increased closer to the O'Shaughnessy mel scale[7] formula performance increases. This increase is, however, not as great as at $\beta = 2.4$. For the second set of experiments the performance was 72.7% at $\beta = 2.0$ and 73.4% at $\beta = 2.4$. The increased performance at this compression $\beta = 2.4$ can be understood in several ways. Firstly the region used in the database is dr1 and this is the one with the fewest number of speakers in the database, and more importantly the gender spread is 24 males as compared to 14 females. Our earlier work [6] and that of [8] shows that spectral compression is sensitive to the gender population and therefore the peak here is most likely the results of the gender population that is used in the test set than absolute performance. Even though in the above cited papers the work was on *speech* recognition rather than *speaker* recognition but it is reasonable to assume that it does have the same impact. The Gaussian kernel SVM used here was not exhaustively evaluated for optimal parameters (σ^2, C), but the results show that spectral compression does affect identification rates and therefore the results confirms the results already obtained for *speech* recognition in literature[6, 8].

[1]This was from a discussion I had with Douglas Reynolds of MIT.

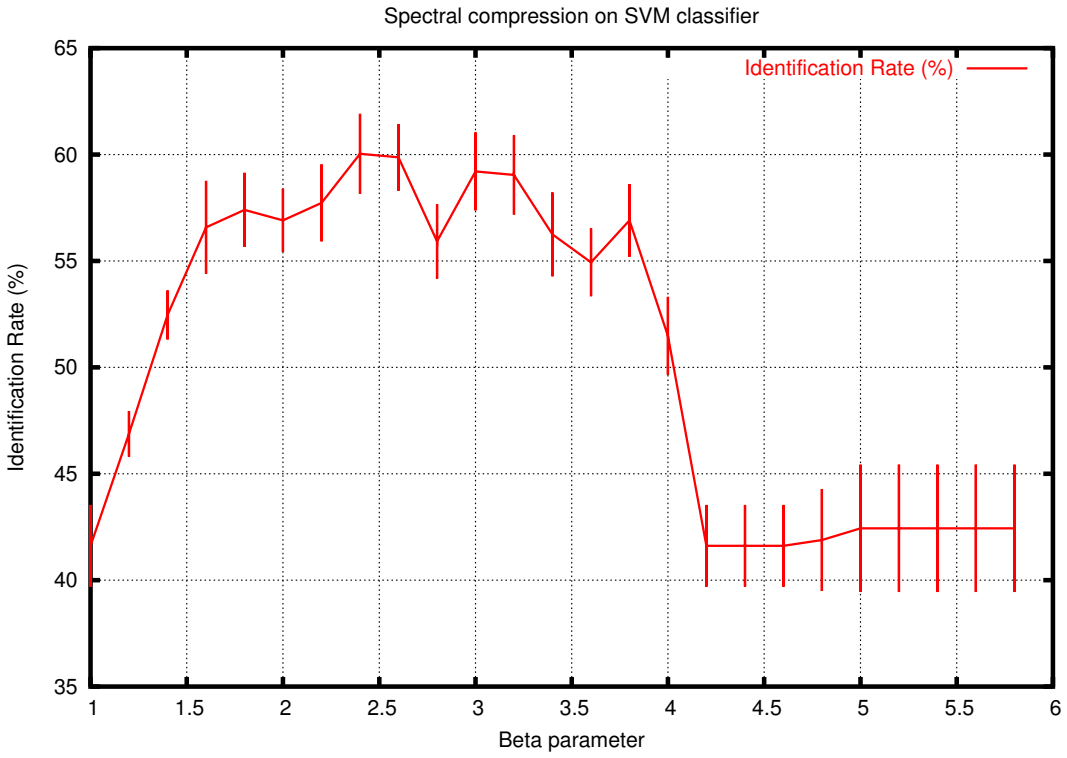

Fig. 5. Identification rate as a function of β with $\alpha = 4$.

5. CONCLUSIONS

The following conclusions can be drawn on the results. Firstly that spectral compression does improve the performance of the system. Too much spectral compression has the possibility of reducing performance. The surprising result from the experiments is that the best performance is obtained at a higher compression level ($\alpha = 4, \beta = 2.4$) than the mel-scale compression. This is most likely the artifact of the database that was used in the experiments. The most likely reason for best performance at this higher compression level is that the database is about 63% males. Earlier work has shown that features-sets with higher spectral compression suits male speakers better. The major weakness of using the SVM classifier is the computational burden. It seems that it should only be used in a support role as a secondary classifier rather than the main classifier given the computational burden. This is the approach that many researchers are adopting as shown in these papers[1, 2]. However, it is pleasing to see that results in literature on the impact of spectral compression are also confirmed when the SVM classifier is used in speaker identification task as should be expected. It shows that indeed spectral compression is important independent of the classifier used.

6. REFERENCES

[1] Shai Fine. A hybrid gmm/svm approach to speech applications. In *Haifa Winter Workshop on Computer Science and Statistics*. Haifa, Israel, 2001.

[2] Shai Fine, George Saon, and Ramesh A. Gopinath. Digit recognition in noisy environments via a sequential gmm/svm system. In *ICASSP*. citeseer.nj.nec.com/fine02digit.html, 2001.

[3] A. Ganapathiraju, J. Hamaker, and J. Picone. Advances in hybrid svm/hmm speech recognition. In *2003*. www.isip.msstate.edu/publications/conferences/gspx/2003/svms/paper_v2.

[4] Rouhana Jhumka and Daniel Mashao. Comparing svm and gmm on speaker identification system. In *PRASA-2002*, pages 47–50, November 2002.

[5] Quan Le and Samy Bengio. Client dependent gmm-svm models for speaker verification. In *icann2003*. http://www.idiap.ch/ bengio/publications/newbib/pdf/quan_2003_icann.pdf, 2003.

[6] D. Mashao and J. Adcock. Utterance dependent parametric warping for a talker-independent hmm-based recognizer. In *ICASSP-97*, volume 2, pages 1235–1238, 1997.

[7] D. O'Shaughnessy. *Speech Communication - Human and Machine*. Addison-Wesley, New York, 1987.

[8] A. Potamianos and R Rose. On combining frequency warping and spectral shaping in hmm based speech recognition. In *ICASSP-97*, volume 2, pages 1275–1278, 1997.

[9] S. Umesh, L. Cohen, and D. Nelson. Fitting the mel scale. In *ICASSP*, 1999.

[10] V. Vapnik. *The Nature of Statistical Learning theory*. Springer, 1995.