

HIGH-DIMENSIONAL LINEAR REPRESENTATIONS FOR ROBUST SPEECH RECOGNITION

Matthew Ager[†], Zoran Cvetković[‡] and Peter Sollich[†]

Department of Mathematics[†] and Department of Electronic Engineering[‡]
King's College London

ABSTRACT

Phoneme classification is investigated in linear feature domains with the aim of improving the robustness to additive noise. Linear feature domains allow for exact noise adaptation and so should result in more accurate classification than representations involving non-linear processing and dimensionality reduction. We develop a generative framework for phoneme classification using linear features. We first show results for a representation consisting of concatenated frames from the centre of the phoneme, each containing f frames. As no single f is optimal for all phonemes, we further average over models with a range of values of f . Next we improve results by including information from the entire phoneme. In the presence of additive noise, classification in this framework performs better than an analogous PLP classifier, adapted to noise using cepstral mean and variance normalisation, below 18dB SNR.

Index Terms— acoustic waveforms, phoneme, classification, robust, speech recognition

1. INTRODUCTION

Many studies have shown that automatic speech recognition (ASR) systems still lack performance when compared to human listeners in adverse conditions that involve additive noise [1, 2, 3, 4]. The systems can improve performance in those conditions by using additional levels of language and context modelling but this context will be most effective when the accuracy of the underlying phoneme sequence is sufficiently free of errors. Hence, robust phoneme recognition is an important stage of ASR. Front-end feature selection is then an important choice to ensure the best phoneme sequence is predicted. In this paper we want to investigate the performance of front-end features, isolated from the effect of high level context. Phoneme classification is commonly used for this purpose and improvements observed can be expected to extend to other recognition tasks [5].

We are particularly interested in linear feature domains. In those domains, additive noise acts additively and hence noise adaptation of Gaussian mixture models can be performed exactly. The ease of noise adaptation in linear feature domains contrasts with the situation for other commonly used speech representations such as mel-frequency cepstral coefficients (MFCC) and perceptual linear prediction coefficients (PLP) [6] that use non-linear processing. The non-linear functions make exact adaptation to noise impossible in practice. However, in order to use acoustic waveforms and benefit from that exact adaptation we must first resolve a number of issues.

Linear representations have been considered previously by other authors, including Poritz [7] and Ephraim and Roberts [8], leading to a recent proposal by Mesot and Barber [9] to use switching linear dynamical systems (SLDS) to explicitly model speech as a time

series [9]. The SLDS approach exhibited significantly better performance at recognising spoken digits in additive Gaussian noise when compared to standard hidden Markov models (HMMs); however, it is computationally expensive even when approximate inference techniques are used. Turner and Sahani proposed using modulation cascade processes to model natural sounds simultaneously on many time-scales [10], but the application of this approach to ASR remains to be explored. In this paper we do not directly use the time series interpretation and impose no direct temporal constraints on the models. Instead, we investigate the effectiveness of the acoustic waveform front-end for robust phoneme classification using Gaussian mixture models (GMMs), as those models are commonly used in conjunction with HMMs for practical applications.

We develop the fixed duration segment models using GMMs with diagonal covariance matrices from [11] and address the issue that there are no analogues of delta features for acoustic waveforms, by instead considering longer duration segments so as to include the same information used by the delta features. The impact of the segment duration is investigated next and we find that no single segment duration is optimal for all phoneme classes, but by taking an average over the duration, the error rate can be significantly reduced. Finally, we also seek to include information from the entire phoneme by incorporating information from five sectors of the phoneme. Many authors have already considered this problem of mapping the variable duration phoneme segments to a representation with fixed dimensionality, first defining sectors of the phoneme and taking the mean over the frames of each sector [12, 13, 14]. Instead we train separate classifiers for each sector and then combine the corresponding log-likelihoods. When this frame averaging and sector sum are both implemented using a PLP+ Δ + $\Delta\Delta$ front-end, we obtain an error rate of 18.5% in quiet conditions, better than any previously reported results using GMMs trained by maximum likelihood. At all stages we consistently find that PLP+ Δ + $\Delta\Delta$ is the most accurate representation in quiet conditions, with acoustic waveform being more robust to additive noise.

2. CLASSIFICATION

Throughout this paper Gaussian mixture models (GMMs) are used to model phoneme densities, trained using the expectation maximisation (EM) algorithm. The probability density function, $p(x)$, $x \in \mathbb{R}^d$, of a Gaussian mixture model with c components has the following form:

$$p(x) = \sum_{i=1}^c \frac{w_i}{(2\pi)^{\frac{d}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp \left[\frac{-(x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i)}{2} \right] \quad (1)$$

where w_i , μ_i and Σ_i are the weight, mean and covariance matrix of the i^{th} mixture component respectively.

We use GMMs with diagonal covariance matrices as it drastically reduces the number of parameters required. This is a common modelling approximation when training data is sparse. In the case of acoustic waveforms we additionally impose a zero mean constraint for models as a waveform x will be perceived the same as $-x$. Typically the log-likelihood is used for calculations; we denote the log-likelihood of x by $\mathcal{L}(x) = \log(p(x))$. Classification is performed using the following function:

$$\mathcal{A}^L(x) = \arg \max_{k=1, \dots, K} \mathcal{L}^{(k)}(x) + \log(\pi_k) \quad (2)$$

where x can be predicted as belonging to one of K classes. The inclusion above of π_k , the prior probability of class k , means that we are effectively maximising the log-posterior probability of class k given x . The prior probabilities are specified as the relative proportion of each class in the training set.

2.1. Model Average

In general, more variability of the training data can be captured by a GMM with an increased number of components, however, if too many components are used, over-fitting can occur. The best compromise is usually located by cross validation using the classification error on a development set. The result is a single value for the number of components required. We use an alternative approach and take the model average over the number of components, c , here having values in $\mathcal{C} = \{1, 2, 4, 8, 16, 32, 64, 128\}$. This effectively gives a mixture of mixtures [15], where the set is uniformly distributed on a log scale to give a good range of model complexity without including too many of the complex models. We compute the model average log-likelihood for $\mathcal{M}(x)$ as:

$$\mathcal{M}(x) = \log\left(\sum_{c \in \mathcal{C}} u_c \exp(\mathcal{L}_c(x))\right) \quad (3)$$

with the model weights $u_c = \frac{1}{|\mathcal{C}|}$.

Alternatively the mixture weights allocated to each model can be determined from the posterior densities of the models on a development set to give a class dependent weighting, i.e.

$$u_c = \frac{\sum_{x \in \mathcal{D}} \exp(\mathcal{L}_c(x))}{\sum_{d \in \mathcal{C}} \sum_{x \in \mathcal{D}} \exp(\mathcal{L}_d(x))} \quad (4)$$

where \mathcal{D} is a development set. Preliminary experiments suggested that using those posterior weights only gives a slight improvement over (3). We therefore choose to take those uniform weights ($u_c = \frac{1}{|\mathcal{C}|}$) for all results shown in this paper.

Figure 1 shows the error rate of the GMM classifiers in quiet conditions as a function of the number of mixture components. Curves are shown for the five representations; acoustic waveforms, PLP, MFCC, PLP+ $\Delta+\Delta\Delta$ and MFCC+ $\Delta+\Delta\Delta$. The solid curves show the error rate for the individual models, dashed curves represent the uniform model average as in (3) up to the number of components on the abscissa. The best results are obtained with $\mathcal{C} = \{1, 2, 4, 8, 16, 32, 64, 128\}$ and $u_c = \frac{1}{8}$ in this case for all representations. We found that the uniform mixture gave very similar results to those derived from the posterior probabilities of development data (3). This is supported by observations in [15] where the authors also took a uniform weight for each of the models. Model averaging gives an improvement of 1.6%, 2.8% and 4.4% for each of the respective representations when up to 128 components are included.

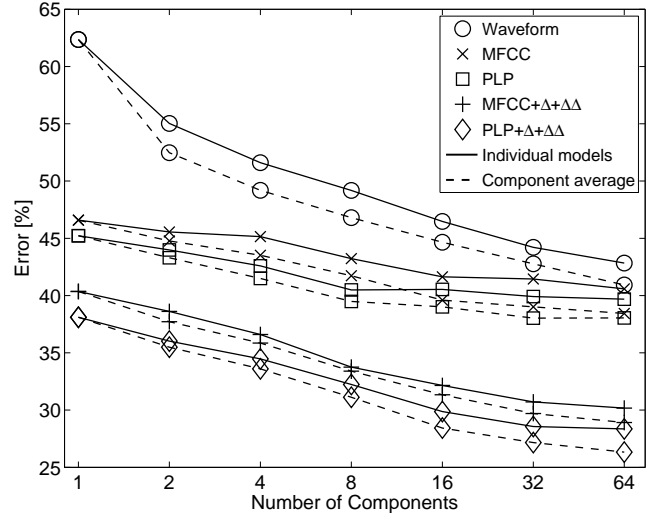


Fig. 1. Model averaging for acoustic waveforms, MFCC and PLP models, all trained and tested in quiet conditions. Solid: GMMs with number of components shown; dashed: average over models up to number of components shown. The model average reduces the error rate in all cases.

2.2. Segment Duration

Ideally all relevant information should be retained by our phoneme representation, but as it is difficult to determine exactly which information is relevant we initially choose to take f consecutive frames closest to the centre of each phoneme and concatenate them. Whilst the precise number of frames required for accurate classification could in principle be inferred from the statistics of the phoneme segment durations, we see in Table 1 that those durations not only vary significantly between classes but also that the standard deviation within each class is at least 24ms. Therefore no single duration can be suitable for all classes. The determination of an optimal f from the data statistics would be even more complicated when $\Delta+\Delta\Delta$ are included, because these incorporate additional information about the dynamics of the signal outside the f frames.

Assuming that no single value of f will be optimal for all phoneme classes we instead consider the sum of the mixture log-likelihoods \mathcal{M}_f , as defined in (3) but now indexed by the number of frames used. The sum is taken over the set \mathcal{F} which contains the values of f with the lowest corresponding error rate, for example $\mathcal{F} = \{7, 9, 11, 13, 15\}$ for PLP:

$$\mathcal{R}(\bar{x}) = \sum_{f \in \mathcal{F}} \mathcal{M}_f(x^f) \quad (5)$$

where $\bar{x} = \{x^f | f \in \mathcal{F}\}$, with x^f being the vector with f frames. Note that we are adding the log-likelihoods for different f , which amounts to assuming independence between the different x^f in \bar{x} . Clearly this is an imperfect model, as e.g. all components of x^7 are also contained in x^{11} and so are fully correlated, but our experiments show that it is useful in practice. Consistent with the independence assumption, in noise we adapt (see Section 2.4 below) the models \mathcal{M}_f separately and then combine them as above. The same applies to the further combinations discussed next.

Table 1. Phoneme duration [ms] in the training data grouped by broad phonetic class.

Group	Min.	Mean \pm std.	Max.
Vowels	2.2	86.0 \pm 46.7	438.6
Nasals	7.6	54.5 \pm 25.6	260.6
Strong Fricatives	14.9	99.5 \pm 38.9	381.2
Weak Fricatives	4.5	68.2 \pm 37.3	310.0
Stops	2.9	39.3 \pm 24.0	193.8
Silence	2.0	94.9 \pm 107.5	2396.6
All	2.0	79.4 \pm 63.4	2396.6

2.3. Sector Sum

We now establish a method to map the variable duration phoneme segments to a fixed length representation for classification. In the previous subsection only frames from the centre of the phoneme segments were used to represent a phoneme. We extend that centre-only concatenation to use information from the entire segment by taking f frames with centres closest to each of the time instants A,B,C,D and E that are distributed along the duration of the phoneme as shown in Figure 2. In this manner the representation consists of five sequences of f frames per phoneme. Those sets of frames are then concatenated to give five vectors x_A, x_B, x_C, x_D and x_E . Models are trained on those five sectors and then the information they provide about each sector is combined, again assuming independence by taking the sum of the log-likelihoods of the sectors:

$$\mathcal{S}(\hat{x}) = \sum_{s \in \{A,B,C,D,E\}} \mathcal{M}_s(x_s) \quad (6)$$

where $\hat{x} = \{x_A, x_B, x_C, x_D, x_E\}$ and \mathcal{M}_s now denotes the model for sector s , using some fixed number of frames f . Both improvements can be combined by taking the sum of the f -averaged log-likelihoods, $\mathcal{R}_s(\bar{x}_s)$, over the five sectors s :

$$\mathcal{T}(\hat{x}) = \sum_{s \in \{A,B,C,D,E\}} \mathcal{R}_s(\bar{x}_s) \quad (7)$$

where $\bar{x}_s = \{x_s^f | f \in \mathcal{F}\}$ with x_s^f being the vector with f frames centred on sector s , and \hat{x} gathers all \bar{x}_s . Given the functions derived above, the class of a test point can be predicted using one of the following:

$$\mathcal{A}_f^M(x) = \arg \max_{k=1,\dots,K} \mathcal{M}_f^{(k)}(x) + \log(\pi_k) \quad (8)$$

$$\mathcal{A}^R(\bar{x}) = \arg \max_{k=1,\dots,K} \mathcal{R}^{(k)}(\bar{x}) + \log(\pi_k) \quad (9)$$

$$\mathcal{A}_f^S(\hat{x}) = \arg \max_{k=1,\dots,K} \mathcal{S}_f^{(k)}(\hat{x}) + \log(\pi_k) \quad (10)$$

$$\mathcal{A}^T(\hat{x}) = \arg \max_{k=1,\dots,K} \mathcal{T}^{(k)}(\hat{x}) + \log(\pi_k) \quad (11)$$

where π_k is the prior probability of predicting class k as in (2).

2.4. Noise Adaptation

As discussed in the introduction, we are interested in using acoustic waveform domains or more generally feature domains where additive noise acts additively. In this paper we assume stationary Gaus-

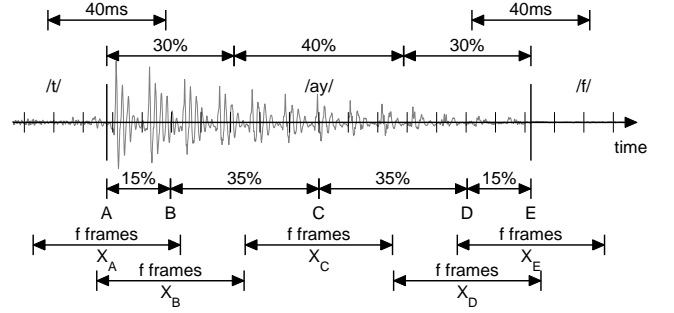


Fig. 2. Comparison of existing phoneme representations. Top: Division described in [12] resulting in five sectors, three covering the duration of the phoneme and two of 40ms over the transitions. Bottom: f frames closest to the five points A,B,C,D and E (that correspond to the centres of the regions above) are selected to map the phoneme segment to five feature vectors x_A, x_B, x_C, x_D and x_E .

sian noise of known variance, where the acoustic waveform classifiers can be adapted for noise corrupted data by modifying all covariance matrices to $\tilde{\Sigma}$:

$$\tilde{\Sigma} = \frac{\Sigma + \sigma^2 \mathbf{N}}{1 + \sigma^2} \quad (12)$$

with \mathbf{N} being the normalised covariance matrix of the noise, in this case estimated from samples of the noise, and σ^2 is the noise variance. A diagonal approximation for \mathbf{N} is used to keep $\tilde{\Sigma}$ diagonal. This approach is reasonable because we work not with the acoustic waveforms directly but with DCT transforms (see below) that approximately decorrelate different feature components.

Exact model adaptation for PLP features is not possible so instead we apply the feature standardisation technique of cepstral mean and variance normalisation (CMVN) which reduces the global effect of the noise on the corrupted feature distribution. We also consider the matched condition classifier that is trained data from conditions that exactly match the test conditions. This is taken as an optimal baseline as other work suggests that matched condition training is superior to any other approach [16].

3. EXPERIMENTS

Realisations of phonemes were extracted from the SI and SX sentences of the TIMIT database. The training set consists of 3,696 sentences sampled at 16kHz. Each sentence is normalised to have on average unit energy per sample. Noisy data is generated by applying additive pink noise extracted from NOISEX-92 at nine SNRs followed by the same average unit energy per sample normalisation. The SNRs were set at the sentence level hence the local SNR of the individual phonemes may differ significantly, causing SNR mismatch at phoneme level. In total there were ten test conditions: -18dB to 30dB in 6dB increments and quiet (Q).

Each sentence was divided into a sequence of 10ms non-overlapping frames to give the acoustic waveform representation. The frames are individually processed using a DCT. Here the DCT is used to decorrelate the waveform frames to improve the diagonal approximation of the GMMs. It is nothing more than an orthogonal transformation and would be an unnecessary stage if full covariance models were used. The processing results in a sequence of 160-dimensional vectors. For comparison, the normalised sentences are

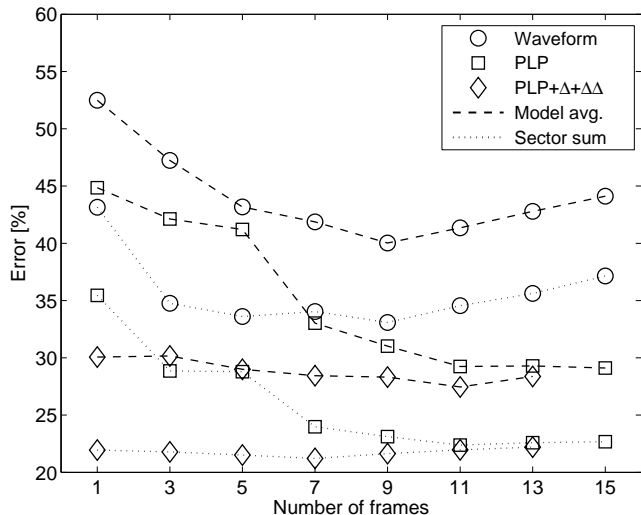


Fig. 3. Error rates of the different representations tested in quiet condition, showing the improvement of the sector sum over the model average as a function of f , the number of frames from each sector.

processed using a standard implementation of PLP. Again frames are extracted every 10ms but 25ms long. The dimensions of the vector sequences are 13 and 39 for PLP and PLP+ $\Delta+\Delta\Delta$ respectively.

Following the extraction of the phonemes and removal of the glottal closures there are a total of 140,225 phoneme realisations. The remaining classes are combined into 48 groups in accordance with [13, 17]. Even after this combination some of the resulting groups have too few realisations. The smallest groups with fewer than 1,500 realisations were increased in size by the addition of temporally shifted versions of the data. All classification tests were carried out on the core test set which is comprised of 7,215 examples. The multi-class test error is computed over the 39 groups detailed in [17] where each group contains similar phonemes and confusions among these are then not counted.

4. RESULTS

Figure 3 shows the relationship between the number of frames concatenated from each sector and the error rates obtained. We see that the best results for acoustic waveform classifiers are achieved around nine frames and eleven frames for PLP without deltas. The PLP+ $\Delta+\Delta\Delta$ features are less sensitive to the number of frames with little difference in error from one to thirteen frames. If we consider the best results obtain for PLP without deltas, 21.4% using eleven frames with the best for PLP+ $\Delta+\Delta\Delta$, 18.5% with seven frames, then the performance gap of 2.9% is much smaller than if we compared error rates where both classifiers used the same number of frames. Clearly it is not surprising that fewer PLP+ $\Delta+\Delta\Delta$ frames are required for the same level of performance as the deltas are a direct function of the neighbouring PLP frames. It is still useful to see that in terms of the ultimate performance on this classification task the two error rates with and without deltas are similar. Those results are directly comparable with the GMM baseline results from other studies shown in Table 1. The error rates obtained using the f -average over the five best values of f are 32.1%, 21.4% and 18.5% for acoustic waveforms, PLP and PLP+ $\Delta+\Delta\Delta$ respectively.

Figure 4 compares the performance of the final classifiers, in-

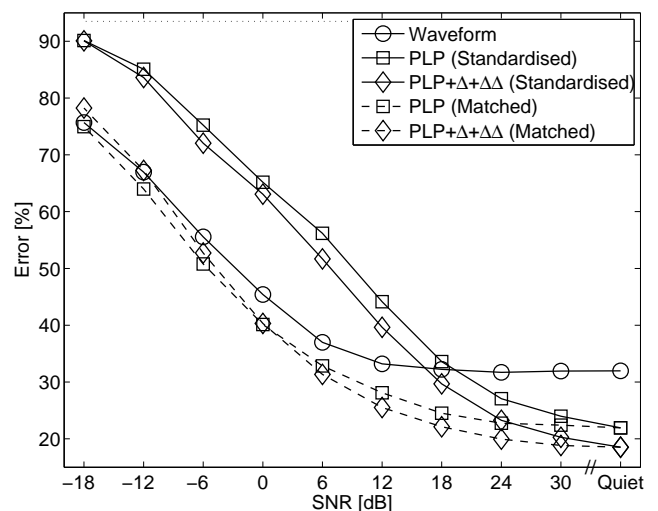


Fig. 4. Performance of the classifiers in pink noise extracted from NOISEX-92. Curves shown for the best representation from Fig. 3 using the f -average. Dotted line indicates chance level at 93.5%.

cluding both the f -average and the sector sum, on data corrupted by pink noise. The solid curves give the results for the acoustic waveform classifier adapted to noise using (12), and for the PLP classifier with and without $\Delta+\Delta\Delta$ trained in quiet conditions and adapted to noise by CMVN. PLP+ $\Delta+\Delta\Delta$ remains the better representation for very low noise, but waveforms give lower errors beyond a crossover point around 18dB SNR, the precise value depending on whether we compare to PLP or PLP+ $\Delta+\Delta\Delta$. As before, they also perform better than chance down to -18 dB SNR. The dashed lines in Figure 4 show for comparison the performance of PLP classifiers trained in matched conditions. As explained, the CMVN and matched curves for PLP provide the extremes between which we would expect a PLP classifier to perform if model adaption analogous to that used with the acoustic waveforms was possible, or some other method to improve robustness was employed such as the ETSI advanced front-end (AFE) [18]). As expected, the matched conditions PLP+ $\Delta+\Delta\Delta$ classifier has the best performance for all SNR. However, in noise the adapted acoustic waveform classifier is significantly closer to matched PLP+ $\Delta+\Delta\Delta$ than PLP+ $\Delta+\Delta\Delta$ with CMVN.

Table 2 shows the absolute percentage error reduction for each of the four classifiers (8)–(11) in quiet conditions, compared to the GMM with the single best number of mixture components and number of frames f . The relative benefits of the f -average and the sector sum are clear. The sector sum gives the bigger improvements on its own in all cases compared to only the f -average, but the combination of the two methods is better still throughout. The same qualitative trend holds true in noise.

Table 2. Absolute reduction in percentage error for each of the classifiers (8)–(11) in quiet conditions.

Model	Waveform	PLP	PLP+ $\Delta+\Delta\Delta$
Model average (\mathcal{A}^M)	1.6	2.8	4.4
f -average (\mathcal{A}^R)	5.6	6.0	6.3
Sector sum (\mathcal{A}^S)	6.7	8.4	8.7
f -average + sector (\mathcal{A}^T)	9.9	10.0	10.4

5. CONCLUSION AND DISCUSSION

In this paper we have studied some of the potential benefits of phoneme classification in linear feature domains directly related to the acoustic waveform, with the aim of implementing exact noise adaptation of the resulting density models. The results can be directly compared to the existing results in Table 3. We used the standard approximation of diagonal covariance matrices to reduce the number of parameters required to specify the GMMs. The issue of selecting the number of components in the mixture models was approached by taking the model average with respect to the number of components for a sufficiently large set of values. This motivated us to further improve the classifiers by using multiple segment durations and then taking the sum of the log-likelihoods. Information from the whole phoneme was included by repeating the process centred at five points in the phoneme.

We would seek to further improve the results by incorporating techniques used by other authors, in particular the use of committee classifiers to combined a number of representations with different parameters. Additionally a hierarchical classification could be implemented to reduce broad phoneme class confusions [5, 19, 20]. There is also scope for further tuning within the method presented by weighting the sector sum and frame average, or allowing the number of frames to be different for each sector.

Table 3. Existing error rates obtained in other studies for a range of classification methods on the TIMIT core test set. Results in this paper are most comparable to the GMM baselines.

Method	Error [%]
GMM baseline [12]	26.3
GMM baseline [19]	24.1
GMM baseline [14]	23.4
GMM (f-average + sector) PLP+Δ+$\Delta\Delta$	18.5
SVM, 5th order polynomial kernel [12]	22.4
Large margin GMM (LMGMM) [13]	21.1
Regularized least squares [14]	20.9
Hidden conditional random fields [21]	20.8
Hierarchical LMGMM H(2,4) [19]	18.7
Committee hierarchical LMGMM H(2,4) [19]	16.7

6. REFERENCES

- [1] G. Miller and P. Nicely, "An Analysis of Perceptual Confusions among some English Consonants," *Journal of the Acoustical Society of America*, vol. 27, pp. 338–352, 1955.
- [2] R. Lippmann, "Speech Recognition by Machines and Humans," *Speech Communication*, vol. 22, no. 1, pp. 1–15, 1997.
- [3] J. Sroka and L. Braidia, "Human and Machine Consonant Recognition," *Speech Communication*, vol. 45, no. 4, pp. 401–423, 2005.
- [4] S. Phatak and J. Allen, "Syllable Confusions in Speech-Weighted Noise," *Journal of the Acoustical Society of America*, vol. 121, no. 4, pp. 2312–2326, 2007.
- [5] A. Halberstadt and J. Glass, "Heterogeneous measurements and multiple classifiers for speech recognition," in *International Conference on Spoken Language Processing*, 1998.
- [6] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.
- [7] A. Poritz, "Linear Predictive Hidden Markov Models and the Speech Signal," in *Proceedings of the International Conference on Acoustic, Speech, and Signal Processing*, 1982, pp. 1291–1294.
- [8] Y. Ephraim and W. Roberts, "Revisiting Autoregressive Hidden Markov Modeling of Speech Signals," *IEEE Signal Processing Letters*, vol. 12, no. 2, pp. 166–169, Feb. 2005.
- [9] B. Mesot and D. Barber, "Switching Linear Dynamical Systems for Noise Robust Speech Recognition," *IEEE Transactions of Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1850–1858, 2007.
- [10] R. E. Turner and M. Sahani, "Modeling Natural Sounds with Modulation Cascade Processes," in *Advances in Neural Information Processing Systems*, 2008, vol. 20.
- [11] M. Ager, Z. Cvetković, and P. Sollich, "Robust Phoneme Classification: Exploiting The Adaptability of Acoustic Waveform Models," in *Proceedings of EUSIPCO*, 2009.
- [12] P. Clarkson and P. Moreno, "On the Use of Support Vector Machines for Phonetic Classification," in *Proceedings of the International Conference on Acoustic, Speech, and Signal Processing*, 1999, vol. 2, pp. 585–588.
- [13] F. Sha and L. Saul, "Large Margin Gaussian Mixture Modeling for Phonetic Classification and Recognition," in *Proceedings of the International Conference on Acoustic, Speech, and Signal Processing*, 2006.
- [14] R. Rifkin, K. Schutte, M. Saad, J. Bouvrie, and J. Glass, "Noise Robust Phonetic Classification with Linear Regularized Least Squares and Second-Order Features," in *Proceedings of the International Conference on Acoustic, Speech, and Signal Processing*, 2007, pp. IV–881–IV–884.
- [15] S. Srivastava, M. Gupta, and B. Frigiyik, "Bayesian Quadratic Discriminant Analysis," *Journal of Machine Learning Research*, vol. 8, pp. 1287–1314, 2007.
- [16] M. Gales and S. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 352–359, 1996.
- [17] K.-F. Lee and H.-W. Hon, "Speaker-Independent Phone Recognition using Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.
- [18] "ETSI document - ES 202 050 - STQ: DSR. Distributed speech recognition; Advance front-end feature extraction algorithm;Compression algorithms;," 2007.
- [19] H. Chang and J. Glass, "Hierarchical Large-Margin Gaussian Mixture Models For Phonetic Classification," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, 2007, pp. 272–275.
- [20] F. Pernkopf, T. Pham, and J. Bilmes, "Broad Phonetic Classification using Discriminative Bayesian Networks," *Speech Communication*, vol. 51, no. 2, pp. 151–166, 2009.
- [21] D. Yu, L. Deng, and A. Acero, "Hidden Conditional Random Field with Distribution Constraints for Phone Classification," in *Proceedings of Interspeech*, 2009.