LDA Based Modelling of Foreign Accents in Continuous Speech

Karsten Kumpf
Speech Technology Research Group
Department of Electrical Engineering
The University of Sydney, Australia
E-mail: karsten@speech.su.oz.au

A foreign accent classification system based on phoneme-dependent LDA models has been implemented. The classifier generates accent likelihood scores for single phoneme segments extracted from continuous speech. An automatic training and model optimisation procedure allows evaluation of the contribution of individual phoneme classes and features to the classification task. The average accent classification rates for single phonemes from three accented speaker groups were 69.4% and 49.5% in a multi-speaker and a speaker-independent test, respectively. The relative positions of the accented speakers in the feature space can be shown.

1. INTRODUCTION

This paper describes a new approach to the foreign accent classification problem in continuous speech. Previously the author reported on an automatic accent classification system that discriminates native Australian English speakers from speakers with Lebanese Arabic and South Vietnamese accents based on likelihood scores from accent-dependent HMM phoneme recognizers (Kumpf and King 1996). In the current experiments, acoustic, prosodic and contextual features are combined for the discriminative training of accent classification models at the phoneme level. The ultimate aim is to improve the robustness and accuracy of the automatic accent classification system.

Recently a number of researchers have focused on the classification of speaker accents. Zissman et. al. (1996) reported on the use of phoneme n-gram language models for the classification of two regional speaker dialects in continuous Latin American Spanish. Arslan and Hansen (1996a, 1996b) applied accent dependent HMM codebooks to the discrimination of single and connected words pronounced by native American English speakers and three foreign accented speaker groups.

In our previous experiments the features extracted from manually and automatically labelled phoneme segments were used to train accent-specific phoneme recognizers. This study concentrates on the modelling of speaker accents for individual phoneme classes using linear discriminant functions. The phonemic segmentation is provided by using either hand labelled or automatically segmented read speech. Linear Discriminant Analysis (LDA) allows the utilisation of complex feature sets and the optimisation of the phoneme-dependent accent discrimination models while providing maximal class separation. The contribution of specific features and individual phoneme classes to the speaker accent classification task can be evaluated. In addition a low-dimensional representation of the speaker positions in the feature space relatively to each other can be obtained.

The design of the classification system and the application of LDA to the modelling of differences between accented speaker groups are outlined in Section 2. Section 3 describes the accented database, the pre-processing and the composition of the feature space. Experimental results are reported in section 4.

2. SYSTEM DESCRIPTION

This section gives an overview over the architecture of the accent classification system and describes the algorithms used for the training of the phoneme-dependent LDA models.

2.1 Algorithm and System Architecture

The foreign accent classification system is designed to be modular and transparent. The processing of the speech utterances and the derivation of the decision over the speaker accent follow a stepwise procedure as outlined in Figure 1. The accent classification system operates on single phoneme segments and the segmentation of the accented training database is obtained either manually or automatically. From each of the phoneme segments one feature vector comprising acoustic, prosodic and contextual information is extracted as outlined in section 3. The classifier is trained by estimation

of accent discrimination LDA models for the segments of each individual phoneme class in the training database.

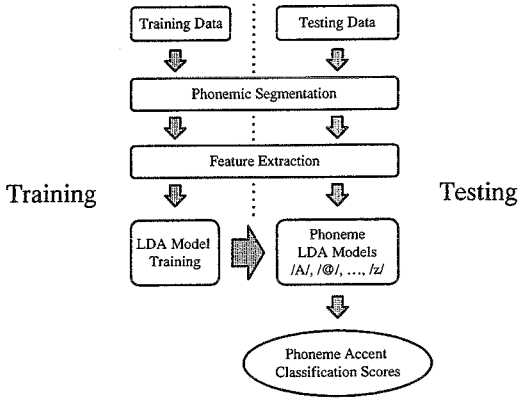

Figure 1: System overview

The phoneme LDA models can be individually optimised to remove features that do not contribute to the accent class separation and to eliminate outliers from the training database.

The utterances of the test database are passed through a phoneme recognizer (approximated by an automatic phoneme labelling system) to provide a rough phonemic segmentation. From each test utterance a sequence of feature vectors $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ is extracted with N being the number of phoneme segments in the utterance. These feature vectors are subsequently processed by their corresponding phoneme LDA models giving a sequence of accent likelihood scores for the phoneme segments. The likelihood scores can be accumulated over the utterance with optional weighting to produce accent likelihood scores for the utterance. The implementation of this step has not yet been finished. The results of the experiments presented in this study focus on the accent class discrimination for single phoneme segments.

2.2 Application of LDA to Accent Discrimination

Linear Discriminant Analysis (LDA) is a fast and robust multivariate discrimination technique that finds linear combinations of the input variables which maximise the class separation by increasing the ratio of between-class to within-class variance (Flury and Riedwyl 1988, Venables and Ripley 1994). The linear discriminants produced by the LDA are uncorrelated and correspond to the eigenvalues of the between-class covariance matrix. The number of linear discriminants is limited to min(p, g-1), with p being the number of input variables and g being the number of classes.

The assumptions on the LDA training data are multivariate normal distribution of the feature space and common covariance structure across all classes. In the accented database each feature vector represents one single phoneme segment. For the data matrix $\mathbf{X}_i = \{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iM}\}$ containing the feature vectors $\mathbf{x}_{im}$ from M phoneme segments of class i the LDA derives $k = 2$ linear discriminant functions $\mathbf{A}_i^k = \{a_{i1}^k, a_{i2}^k, \ldots, a_{ip}^k\}$. This transformation in the two-dimensional space allows to obtain scatter plots for the visualisation of the class separation. The three accent classes A_c are assumed to have equal a-priori probabilities π_c. For each feature vector $\mathbf{x}$ the accent class likelihoods $P(A_c|\mathbf{x}) = \frac{P(\mathbf{x}|A_c) \cdot \pi_c}{P(\mathbf{x})}$ can then be estimated through the Mahalanobis distance of the feature vectors to the class means μ_{ic} in the discriminant variable space. For the derivation of the single phoneme accent classification rates

presented in this paper a maximum likelihood criterion is applied to the decide over the speaker accent for each phoneme segment.

$$A_c = \arg\max_c \left[P(A_C|\mathbf{x})\right] = \arg\max_c \left[-\frac{1}{2}(\mathbf{x}-\mu_{ic})^T \Sigma^{-1}(\mathbf{x}-\mu_{ic}) + \log \pi_c\right], \qquad A_c \in \{\text{AuE}, \text{LA}, \text{SV}\}.$$

2.3 LDA training and model optimisation

A stepwise automatic procedure is applied to estimate an optimised LDA model for each phoneme class. The parsimonious model uses the smallest possible number of feature variables while ensuring maximal accent class separation. This reduces the computational complexity and allows a ranking of the features due to their contribution to the accent discrimination task.

For each LDA phoneme model redundant feature variables in the training set are first eliminated based on their standard deviation and their multiple correlation with the other features, in order to improve the numerical stability of the LDA. Three different methods have been developed and tested for the subsequent stepwise model optimisation. The first technique continues to rank the feature variables based on their multiple correlation with each other. The second method drops the features based on their significance in the regression of the variables on the LDA scores. The third approach performs a multivariate analysis of variance of the feature variables on all LDA vectors simultaneously. All optimisation techniques estimate a very similar ordering of the feature variables. The simple regression method provides the best model optimisation and behaves most stable for small training data sets.

3. DATABASE AND PREPROCESSING

The speech corpus used in this work is the part of the Australian National Database of Spoken Language. This accent class experiment has been confined to three accented male speakers groups with native Australian English, Lebanese Arabic and South Vietnamese accents. Table 1 outlines the size of the corpus available for system training and testing.

	AuE	LA	SV
No. of speakers	22	26	24
No. of utterances	3650	1450	1350
No. of phoneme segments	158166	61602	55887

Table 1: Summary of accented database

3.1 Acoustic and Prosodic Features

The usefulness of acoustic features extracted from continuous speech for the discrimination of speaker accents and languages has been demonstrated by various researchers (Arslan and Hansen 1996, Zissman et. al. 1996, Kumpf and King 1996). In this study the acoustic feature vector consists of the first 12 MFCC coefficients and log energy which are extracted from a 16 ms analysis window around the centre of each phoneme segment.

Linguistic and phonetic theory suggests that some information about the speaker accent can be captured with prosodic features such as duration, pitch and stress. Currently phoneme segment duration is included in the multivariate feature set and the use of pitch tracks is intended.

3.2 Contextual Features

In order to capture the accent-specific pronunciation differences of the phoneme segments in their phonetic context, the feature vector of each segment is enriched with contextual information. The phoneme string delivered by the manual or automatic segmentation of the speech utterances allows the construction of a categorical feature set describing the left and right phonetic context of each phoneme segment. The contextual feature set includes the broad phonetic class of the context (vowels, diphthongs, consonants and pauses) and some phonetic features related to that phonetic class as outlined in Table 2.

The categorical context features are coded in a 50-dimensional numerical feature vector using Helmert contrasts for inclusion in the LDA training algorithm. The combination of the acoustic-prosodic and contextual features thus expands a 65-dimensional feature space. Each data point in this feature space represents a phoneme segment in the database.

Contextual Feature		Levels
broad phonetic class		4
voicing		3
consonant features	type of articulation	6
	place of articulation	7
vowel features	vowel height	4
	vowel position	4
	vowel rounding	3

Table 2: Summary of categorical contextual features

4. EXPERIMENTAL RESULTS

In order to test the accented speaker group separation based on phoneme-dependent LDA models a multi-speaker test was performed. The manually labelled utterances of 22 speakers from the three accented speaker groups were used for system training and testing. Table 3 summarises the classification results. Initially a 14-dimensional feature variable space containing only acoustic features and phoneme segment duration was used for the LDA training (row 1 of table 3). The average phoneme-accent classification rate on single phoneme segments is 64% across all phoneme classes.

The inclusion of the contextual feature variables yielded an performance increase of about 7.8% and expanded the feature space to 65 dimensions, although on average only 48 features remained in LDA models after the removal of numerically redundant variables (row 2 of table 3). The best accent classification was achieved for the phoneme segments of long vowels and diphthongs, while accents of the stop consonants and /schwa/ were the most difficult to discriminate.

	Classification (% correct)				Average
	AuE	LA	SV	Average	No. of features
1) MFCC, duration and log E features	68.7	63.4	60.1	64.1	14
2) all non-redundant features	74.7	66.7	66.3	69.2	48
3) maximum performance feature set	74.7	67.2	66.5	69.4	38
4) 98% of maximum performance	73.7	65.9	65.2	68.3	25
5) features set reduced to 14 variables	71.9	63.2	62.3	65.7	14

Table 3: Accent classification rates and feature set sizes for LDA model training

A further tuning of the models to maximum performance with the optimisation techniques outlined in section 2.3. reduced the average number of features required to 38 (row 3 of table 3) while giving the accent highest classification rate. With a loss of less than 2% in the overall classification rate the feature space could be further reduced to 25 dimensions (row 4 of table 3). Finally, a further reduction of the feature space to 14 variables per phoneme LDA model resulted in an accent discrimination rate slightly higher than that achieved when using the acoustic and prosodic features only (row 5 of table 3).

Figure 2a shows the accent class separation for the phoneme segments of a single phoneme class in the linear discriminant variable space. The accent class means are clearly separated. The discriminant scores of the segments belonging to each speaker were averaged to show the positions of the accented speakers in the discriminant variable space relatively to each other (Figure 2b).

The stepwise LDA model feature space reduction revealed the contribution of each feature variable to the accent classification task. The advantage of tuning the models individually became evident, as the

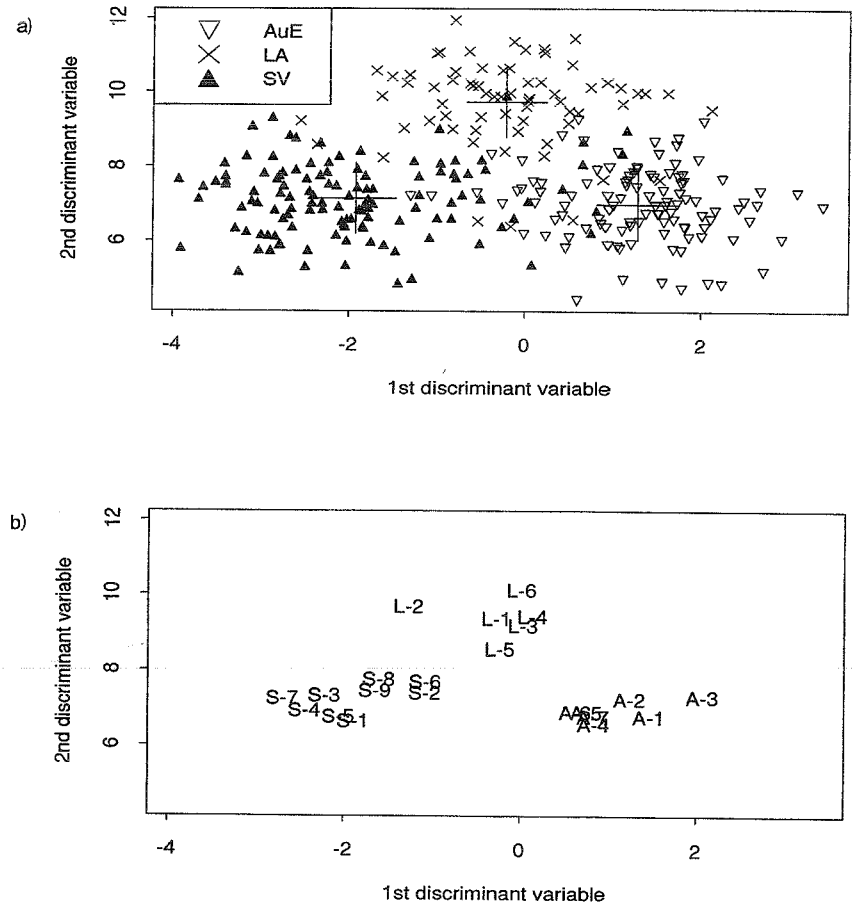

Figure 2: a) Separation of segments of phoneme class /e:/ in the LDA space
b) Speaker group separation based on phoneme segments of class /e:/.

composition and ranking of the optimal feature set was specific to each phoneme class. Most of the MFCC coefficients were retained in all optimised LDA models. The energy coefficient was also included in most models but was especially highly ranked for the unvoiced stops and fricatives. Phoneme segment duration was important in almost all phoneme LDA models except for /u:/, /@u/ and /Z/.

The feature describing the phonetic class of the left context of the phoneme segments was highly ranked for short vowels and stop consonants. The phonetic class of the right context was most important for /r/, /d/, /dZ/ and /tZ/ segments and also highly ranked for the vowel and diphthong classes. The features coding the type and place of articulation of consonants in the context of the phoneme segments were used in the LDA models for most vowel and diphthong phoneme classes. By contrast the features describing vowel height, position and rounding appeared mostly when characterising the context of consonant segments. The vowel feature rounding seemed to be of little importance and was only used for the phoneme LDA models of /S/, /k/, /r/ and /v/.

In a final test the performance of the phoneme LDA models was evaluated in a speaker-independent accent classification test on the automatically labelled phoneme segments of the remaining 52 speakers in the database (table 4).

Classification (% correct)				Average
AuE	LA	SV	Average	No. of features
57.3	41.9	46.7	49.5	48

Table 4:

The average phoneme-accent classification rate was 49.5%, much lower than in the multi-speaker test. The author concludes that the accented speaker sets used for the accent model training were too small and did not represent the whole speaker groups well enough. Some loss of performance may also be due to systematic differences between the manual labelling of the training data and the automatic segmentation of the test data. For the processing of whole utterances with the automatic accent classification system it will be necessary weight the likelihood scores from the individual phoneme LDA models due to their average classification rates.

5. CONCLUSION

It has been shown that multivariate feature vectors containing acoustic, prosodic and contextual information can be used to generate foreign accent likelihood scores for single phoneme segments. The contribution of individual phoneme classes and of specific feature variables to the accent classification task has been evaluated. The position of the speakers relatively to each other in the feature space could be analysed.

The ongoing work focuses on the inclusion of more features that contain accent dependent information and the development of the accent classification system for continuous speech utterances. Also a human perception experiment has been conducted to establish benchmarks for foreign accent classification.

ACKNOWLEDGEMENTS

I would like to thank Andrew Hunt, Robin King and Chris Cleirigh for many inspiring discussions and their comments on the design and evaluation of the experiments. I also wish to thank Bill Venables for helpful insights into the statistical modelling techniques.

REFERENCES

Arslan, L.. M., Hansen, H. L., *Language Accent Classification in American English*, 1996, Speech Communication 18, 353-367.

Duda, R. O., Hart, P. E., *Pattern Classification and Scene Analysis*, 1973, John Wiley, New York.

Flury, B., Riedwyl, H., *Multivariate Statistics, a Practical Approach*, 1988, Chapman and Hall, New York.

Hunt, A. J., *Models of Prosody and Syntax and their Application to Automatic Speech Recognition*, 1995, Ph.D. Thesis, University of Sydney.

Kumpf, K., King, R. W., *Automatic Accent Classification of Foreign Accented Australian English Speech*, Proceedings ICSLP 1996, forthcoming.

Venables W. N., Ripley, B. D., *Modern Applied Statistics with S-Plus*, 1994, Springer, New York.

Vonwiller, J. P., et. al., *Speaker and Material Selection for the Australian National Database of Spoken Language*, 1996, Journal of Quantitative Linguistics, 27.

Zissman, M. A., et. al., *Automatic Dialect Identification of Extemporaneous, Conversational, Latin American Spanish Speech*, Proceedings ICASSP 1996, 777-780.