Foreign Speaker Accent Classification using Phoneme-Dependent Accent Discrimination Models and Comparisons with Human Perception Benchmarks

Karsten Kumpf and Robin W. King[1]

Speech Technology Research Group
Department of Electrical Engineering, University of Sydney, NSW 2006, Australia
[1]Faculty of Information Technology, University of South Australia, SA 5095, Australia
Email: karsten@speech.usyd.edu.au, robin.king@unisa.edu.au

ABSTRACT

This paper reports on the development of a foreign speaker accent classification system based on phoneme class specific accent discrimination models. This new approach to the problem of automatic accent classification allows fast and reliable prediction of the speaker accents for continuous speech through exploitation of the accent specific information at the phoneme level. The system was trained and evaluated on a corpus representing three speaker groups with native Australian English (AuE), Lebanese Arabic (LA) and South Vietnamese (SV) accents. The speaker accent classification rates achieved by our system come close to the benchmarks set by human listeners.

1. INTRODUCTION

A number of recent studies have shown that the performance of speaker-independent speech recognition systems can be improved by modelling of the speaker variability due to regional dialects and foreign speaker accents (eg. Beattie et. al. [2], Brousseau and Fox [4], Van Compernolle et. al. [8], Arslan and Hansen [1]). Various front-end systems for the automatic identification of speaker accents and dialects have been proposed. Blackburn et. al. [3] used a cascade of Multi-layer Perceptrons for the speech segmentation and subsequent foreign speaker accent classification. Zissman et. al. [9] discriminated dialects of Latin American Spanish by combining phoneme recognition with dialect-dependent language modelling. Other approaches derived decisions based on the likelihood scores produced by accent-dependent HMM phoneme recognizers (Kumpf and King [5], Arslan and Hansen [1]).

The accent classification system presented in this paper extends our recently developed technique (Kumpf [6]) for accent classification using Linear Discriminant Analysis (LDA) on individual phoneme classes to continuous speech utterances of variable length. Accent likelihood scores are generated for each phoneme segment in the utterance and accumulated to produce accent discrimination scores for the utterance.

Operation on the phoneme level was motivated by our interest in the capability of each phoneme class to contribute to the speaker accent discrimination task. The results of our previous experiments have shown that the segments of all phoneme types contain accent specific information which can be exploited by training phoneme class-dependent accent discrimination models [6]. We also found that each extracted feature's contribution to the speaker accent separation depends on the phoneme segment class. An individual feature set optimisation for each phoneme class produces the best performing and most robust accent discrimination model set.

Our speaker accent classification system was designed to be flexible and portable to process other accented databases. Therefore it was not feasible to rely on manual data segmentation. An HMM phoneme segmenter was used to automatically segment the accented speech.

The algorithm and architecture of the speaker accent classification system are described in Section 2. Section 3 introduces the speech corpus used in the experiments. Section 4 outlines a study on the human perception of foreign speaker accents which is used here to provide performance benchmarks. Section 5 presents the experimental accent classification results.

2. CLASSIFICATION METHOD

The speaker accent classification for continuous speech utterances is based on the accumulation of accent likelihood scores produced by phoneme class-dependent accent discrimination models [6]. A single feature vector is extracted from each phoneme segment in the training and test utterances which combines (i) acoustic (12 MFCC coefficients and log energy), (ii) prosodic (phoneme segment duration, in Section 5.2. F0 and delta-F0 are added) and (iii) contextual information (categorical features describing the phonetic left and right context of the segment). The coding of the categorical features with contrasts expands the feature space to 65 dimensions. Silence segments and non-speech sounds are ignored.

The feature vectors of each phoneme class are pooled across the training database and used to estimate phoneme-dependent LDA models (sets of linear discriminant functions that provide maximal separation of the accent classes). An optimised feature sub-set is selected for the training of each LDA model to maximise the accent discrimination performance. The optimisation

scheme eliminates features with very low variance and highly correlated features. The remaining features are ranked according to their accent discrimination capability for the individual phoneme classes. The algorithm drops features one by one from the training set based on their significance in the regression of the linear discriminant functions onto the feature set.

During testing the likelihood score of accent A_c for the feature vector $\mathbf{x}_n$ of each phoneme segment n is given by

$$P(A_c|\mathbf{x}_n) = p(\mathbf{x}_n|A_c)\pi_c / p(\mathbf{x}_n), \quad A_c \in \{\text{AuE, LA, SV}\}$$

The probability densities $p(\mathbf{x}_n|A_c)$ are inversely proportional to the exponential of the Mahalanobis distance between the feature vector and the accent class means in the discriminant variable space. The a priori accent class probabilities π_c are assumed to be equal for all accented speaker groups.

The accent discrimination scores S_c for continuous speech utterances of variable length are derived through accumulation of the phoneme-dependent accent likelihood scores over N phoneme segments in the utterance:

$$S_c = \prod_{n=1}^{N} P(A_c|\mathbf{x}_n)$$

S_c is equivalent to the summation of the distances of the feature vectors to the accent class means in the LDA space. The speaker accent decision is derived by choosing the accent class with the smallest accumulated distance:

$$A = \arg\max_{c} \prod_{n=1}^{N} P(A_c|\mathbf{x}_n)$$

For statistically independent feature vectors S_c is equal to the accent likelihood scores for the utterance and the above equation represents the maximum-likelihood criterion. Due to its simplicity and efficiency, we adopted the same model even though the feature vectors extracted from the phoneme sequence contain explicit contextual information (and are therefore not independent).

Our approach differs from that pursued by Miller and Trischitta [7], who applied LDA to the discrimination of four American English dialects. In their system the features extracted from a selection of between 5 and 17 phoneme classes (mostly vowels) are averaged over 75 to 400 utterances per speaker and combined into a single feature vector. Our approach is more flexible as it allows us to (i) model the phoneme-dependent accent information, (ii) train phoneme class-dependent LDA models which reduces the high dimensional feature space and increases the numerical stability and accuracy of the LDA, (iii) process and score utterances progressively.

3. DATABASE

The speech corpus used in this study is part of the Australian National Database of Spoken Language (ANDOSL) and comprises read speech utterances from 72 male speakers of the three accented speaker groups AuE, LA and SV containing 22, 26 and 24 speakers, respectively. All speakers were at least 17 years old and the migrants had spent at least 4 years in Australia. Each speaker produced either 50 or 200 sentences, which resulted in a total of 6450 utterances with an average utterance length of 4.4 seconds. The utterances were segmented using the 44 phoneme classes of the target language Australian English. Manually segmented data was available for 22 speakers. Automatic segmentation was performed using an HMM phoneme aligner trained on a separate set of AuE speakers alignment and the orthographic transcription of the speech utterances.

4. HUMAN PERCEPTION

This section briefly outlines a human perception study which was conducted on 20 native AuE subjects in order to establish evaluation benchmarks for the automatic accent classification task. Each listener classified the speaker accent type and judged the relative accent strength of a randomly selected set of 144 speech segments of up to six seconds duration from all 72 speakers in the database. Statistical analysis was used to assess the influence of (i) the speaker language background, (ii) the segment duration, (iii) the duration of the experiment on the subjects' perception of speaker accent type and strength. Post-experimental interviews were evaluated to identify clues used by the listeners to distinguish the accents. The subjects confidently, consistently and correctly identified the AuE accents and mainly confused LA speakers with either AuE or SV speakers, depending on the relative accent strength.

5. EXPERIMENTAL RESULTS

The amount of training data was varied during the development of the foreign speaker accent classification system, while all tests were carried out on 4750 automatically segmented utterances from 15 AuE, 20 LA and 15 SV speakers.

5.1. Basic Accent Classification System

This section analyses the effects of speech segmentation quality and the amount of training on the performance of the speaker accent classification system (Table 1).

Training data		% correct utterance classifiation			
Speakers	Segmentation	AuE	LA	SV	Average
22	manual	91.5	64.5	77.6	77.9
49	automatic	89.3	71.4	73.9	78.2
62	automatic	91.7	69.8	83.9	81.8

Table 1: Speaker-independent accent classification for varied traning database segmentation and size

The first accent classification system was trained on 1700 manually labelled utterances from 22 speakers. As a result of the feature set optimisation for maximum accent

discrimination on the phoneme level, on average 40 features were selected for the training of each phoneme-dependent LDA model. The accent classification rate on the utterances of the training set was 99.4%, averaged over all speaker groups. In the speaker-independent test on the utterances of the 50 test speakers the classifier achieved an average accent classification rate of 77.9% (first row of Table 1). The accent classification rate on the single phoneme segments of the same test set was only 49.4%. This underlines the benefit of accumulating the phoneme-dependent accent discrimination scores. In agreement with previous experiments and the judgements by the human listeners the performance is worst for the LA speakers, as they are confused both with AuE and SV speakers.

A second experiment investigated how the reduced quality of the automatic segmentation (substitution and insertion errors as well as the shifting of phoneme boundaries) affects the quality of the accent discrimination models. The accent classification system was trained and tested on automatically aligned speech data from 50 speakers. The feature sets for the phoneme-specific LDA models were the same as before (optimised on manually labelled data). A leaving-one-out training procedure was employed to allow speaker-independent testing, using one speaker at a time for testing and the other 49 for training. The average accent classification rate across all speaker groups was 78.2% (row two of Table 1). This performance is very similar to that achieved by the classifier trained on the manually segmented data, however, the number of speakers and the amount of data used for training had more than doubled. These results show that the lower quality of the automatic segmentation can be compensated by increasing the speaker variety in each accented speaker group and the overall size of the training database.

In order to maximise the generalisation of the phoneme dependent accent discrimination models, the number of speakers for the system training was further increased to 63. Again leaving-one-out training was used and on the 50 speaker test set the accents were classified correctly for 81.8% of the utterances (row three of Table 1), which represents a relative error reduction of 16.5%, compared to the previous system. Our best performing HMM based accent classifier, which combined accent dependent phoneme recognizers with statistical language models [5], had been trained under the same conditions and achieved an average classification rate of 79.2% on the same test set. In addition to the higher performance the LDA based system has the advantage of greatly reduced system complexity. Both systems require similar efforts for speech segmentation and feature extraction but (i) the estimation of linear discriminant functions is computationally less expensive than iterative Baum-Welch re-estimation of the HMM models, (ii) the HMM based system requires separate processing of the test utterances through each accent-dependent phoneme recognizer.

5.2. Pitch modelling

The human perception tests indicated that the listeners based their accent classification decisions on acoustic features of individual speech sounds, the intelligibility of single words and utterance content as well as prosodic features such as pitch movements, rhythm and pausing. Since the rhythmic patterns of some speakers' heavily accented speech are more associated with poor reading skills than with the accent itself, we chose not to model rhythm explicitly, but included only pitch data. The feature sets were augmented with level and slope of the pitch contour, extracted from a 26 ms window around the segment centres. The feature set optimisation algorithm ranked the F0 level as a prominent feature for the accent discrimination on most vowel and voiced consonant classes, while the influence of delta-F0 was negligible. The inclusion of F0 resulted in modest increases of the accent classification rates on single phoneme segments, but lead only to slight performance improvements on the utterance level. A possible explanation is that the model contains insufficient context to model the more significant pitch variation correlates of accent.

5.3. Utterance duration

Figure 1 shows the performance of the final speaker accent classification system of Section 5.2. (including the F0 feature) as a function of utterance duration.

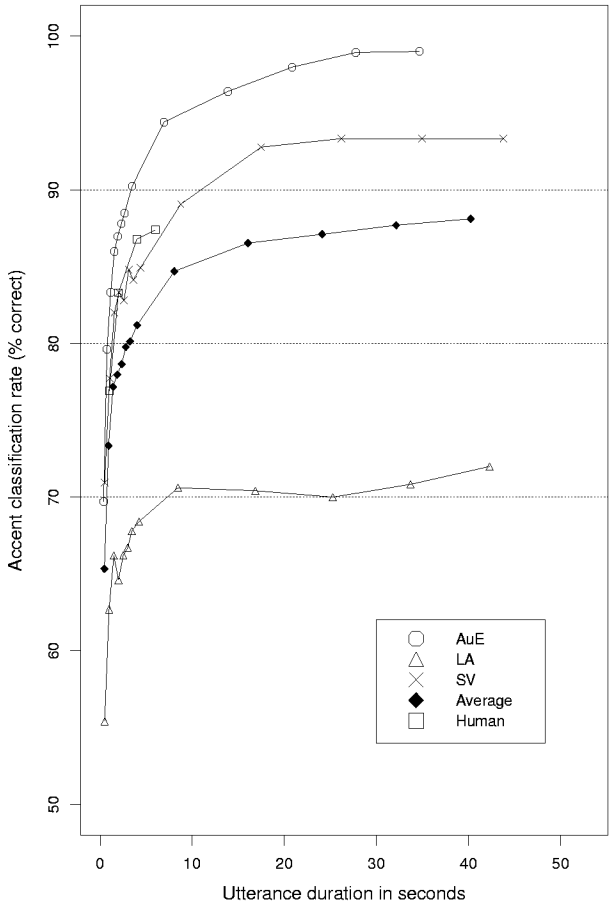

Figure 1: Accent classification rates vs. utterance duration

The duration represents the accumulated length of the phoneme segments processed by the classifier. For some of the migrant speakers the true utterance durations were up to 20% longer due to extensive pausing caused by their problems with the reading of the target language. Up to ten spoken sentences from the same speaker (containing about 425 phoneme segments) were concatenated to simulate the system performance on continuous speech utterances of up to about 40 seconds duration. The tick-marks on the classification curves correspond to 5, 10, 15, 20, 25, 30 and 35 phoneme segments, followed by 1, 2, 4, 6, 8 and 10 concatenated utterances. Due to higher speaking rates the AuE speakers' utterances are shorter in duration, however they contain roughly the same number of phoneme segments. Figure 1 also shows the overall accent classification rate achieved by the human listeners for speech utterances of up to 6 seconds duration.

The average speaker accent classification rate of the automatic system increases rapidly with the accumulation of the phoneme segment accent likelihood scores to 84.7% at 8.1 seconds and reaches 88.1% for the longest durations. The human listeners exploited the accent specific information more efficiently than the automatic classifier and reached an accent classification rate of 87.4% for speech segments of only 6 seconds duration.

5.4. Feature set reduction

Finally, we analysed the influence of the feature set size on the accent classification performance by stepwise reducing the feature set size for the training of the accent discrimination models and thus trading off computational effort against classification accuracy. The tick-marks correspond to 100%, 95% and 90% of the maximum accent classification rate on single phoneme segments.

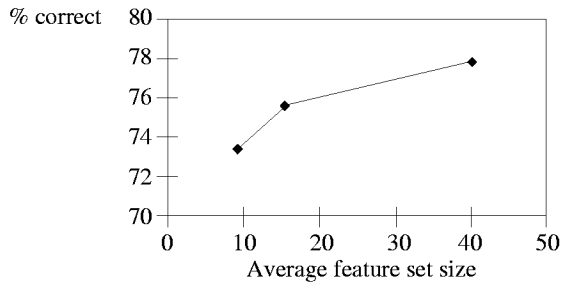

Figure 2: Accent classification rate versus feature set size for utterances (4.4 seconds average duration).

Figure 2 shows that an average reduction of the feature set size from 40 to 9 features results in a 20% increase of the accent classification error (relative) on whole utterances. This underlines the effectiveness of the feature selection algorithm, as most of the relevant information for the speaker accent discrimination is concentrated in the features with the highest ranking.

6. CONCLUSION

We have presented an approach to the complex task of automatic foreign speaker accent classification that successfully exploits the accent specific information extracted from phoneme segments. The new system is more transparent and requires less computational effort than our previous HMM based classifier, combined with an increased capability to discriminate speaker accents. We demonstrated that the classification performance depends on the quality and amount of the training data as well as optimal feature selection. The performance of human listeners on the comparable task is superior to our system, however our human perception study indicates that the listeners have the advantage of combining the processing of low-level features with their morphological and syntactic knowledge of the langua

7. ACKNOWLEDGEMENTS

We thank Andrew Hunt for his valuable suggestions and SPLUS signal processing library and Paul Bagshaw for making available his prosody feature extraction software. We also thank Cyril Latimer for his assistance with the design and evaluation of the human perception study.

8. REFERENCES

[1] L. M. Arslan and J. H. L. Hansen, "*Language Accent Classification in American English*", Speech Communication, Vol. 18, no. 4, pp. 353-367, 1996.

[2] V. Beattie et. al., "*An Integrated Multi-Dialect Speech Recognition System With Optional Speaker Adaptation*", Proceedings EUROSPEECH 1995, Madrid, pp. 1123-1126.

[3] C. S. Blackburn, J. P. Vonwiller, R. W. King, "*Automatic Accent Classification Using Artificial Neural Networks*", Proceedings EUROSPEECH 1993, Berlin, pp. 1241-1244.

[4] J. Brousseau, S. A. Fox, "*Dialect-Dependent Speech Recognizers for Canadian and European French*", Proceedings ICSLP 1992, Banff, pp. 1003-1006.

[5] K. Kumpf and R. W. King, "*Automatic Accent Classification of Foreign Accented Australian English Speech*", Proceedings ICSLP 1996, Philadelphia, pp. 1740-1743.

[6] K. Kumpf, "*LDA Based Modelling of Foreign Accents in Continuous Speech*", Sixth Australian International Conference on Speech Science and Technology, 1996, Adelaide, pp. 257-261.

[7] D. R. Miller and J. Trischitta, "*Statistical Dialect Classification Based on Mean Phonetic Features*", Proceeds. ICSLP 1996, Philadelphia, pp. 2025-2027.

[8] D. Van Compernolle et. al., "*Speaker Clustering for Dialect Robustness in Speaker Independent Recognition*", Proceedings EUROSPEECH 1991, Genova, pp. 723-726.

[9] M. A. Zissman et. al., "Automatic Dialect Identification of Extemporaneous, Conversational, Latin American Spanish Speech", Proceedings ICASSP 1996, Vol. II, pp. 777-780.