AUTOMATIC ACCENT CLASSIFICATION OF FOREIGN ACCENTED AUSTRALIAN ENGLISH SPEECH

Karsten Kumpf [1] *and Robin W. King*

Speech Technology Research Group
Department of Electrical Engineering, The University of Sydney
NSW 2006, Australia
E-mail: karsten@ee.usyd.edu.au, robink@ee.usyd.edu.au

ABSTRACT

An automatic classification system for foreign accents in Australian English speech based on accent dependent parallel phoneme recognition (PPR) has been developed. The classifier is designed to process continuous speech and to discriminate between native Australian English (AuE) speakers and two migrant speaker groups with foreign accents, whose first languages are Lebanese Arabic (LA) and South Vietnamese (SV). The training of the system can be automated and is novel in that it does not require manually labelled accented data. The test utterances are processed in parallel by three (AuE, SV and LA) accent-specific recognizers incorporating the accent-specific HMMs and phoneme bigram language models to produce accent discrimination likelihood scores. The best average accent classification rates were 85.3% and 76.6% for accent pair and three accent class discrimination tasks, respectively. Analyses of the contributions to accent discrimination by the phoneme level processing, and by the language model, are described.

1. INTRODUCTION

This paper describes the development and the experimental performance of an automatic foreign accent identification system for Australian English (AuE) speech. In countries with large migrant populations like Australia there is a wide range of strong accents amongst people whose first language is a foreign language. The speaker variations due to foreign accents complicate the task of automatic continuous speech recognition. Possible solutions for speaker and accent independent speech recognition could be adaptation to the accented speech, processing through accent dependent recognition channels and the utilization of accent specific phonetic and phonological knowledge. The reliable classification of foreign accents is thus a valuable preprocessing step for robust speaker independent speech recognition.

A number of researchers have recently published their work on accent and dialect identification. Blackburn et. al. [1] introduced an algorithm based on speech segmentation and accent classification with MLP which worked well on a small accented speech database. Itahashi and Yamashita [3] undertook extensive studies on the discriminative power of F0 contour based features in combination with PCA and LDA for the identification of Japanese dialects. Hansen and Arslan [2] used HMM codebooks based on acoustic and prosodic features to discriminate between foreign accents in American English. Zissmann [6] reported on the identification of regional American dialects with a language identification system based on phoneme recognition and phonotactic language modelling.

Accent classification differs from language identification in that all speakers are speaking the same target language. However, the speakers with foreign accents are expected to import some of the acoustic and phonological features from their first languages into the speech production process. In the experiments described below we investigated whether (1) the acoustic differences in the production of the AuE phoneme set by the accented speaker groups, and (2) the phonotactic differences in their continuous speech utterances due to phoneme substitutions and approximations, can be exploited for accent discrimination.

An HMM-based automatic segmenter trained on AuE phoneme classes was used to segment the accented speech to create material for subsequent system training. This approach has the merit of not requiring hand labelled accented speech or accent-specific pronunciation dictionaries. The automatically segmented data were then used to train accent-specific HMM phoneme models and to derive accented phoneme bigram language models. For accent classification we employ a maximum likelihood criterion similar to that used by Zissmann and Singer [5] for language identification.

Several multi-speaker and speaker independent accent classification systems were trained and tested in this study in order to discriminate between accented speaker groups. The development of the algorithm and system architecture is described in Section 2. Section 3 reports on the experimental accent classification results, and Section 4 summarizes the system performance, highlights the problems encountered and gives a preview of future work.

2. SYSTEM DESCRIPTION

This section describes the algorithm of the accent classification system, the speech corpus and the feature extraction and processing.

2.1. Algorithm and Architecture

The foreign accent classification system is based on a maximum likelihood criterion applied to the likelihood scores produced by accent-dependent phoneme recognizers that process the unknown utterances in parallel, as shown in Figure 1.

The speech signal is represented by the observation sequence of feature vectors $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T\}$ with T being the number of frames in the utterance. For each accent dependent recognizer a

[1]Karsten Kumpf was supported by a scholarship from the German Academic Exchange Service DAAD.

phoneme HMM set Λ_A and a language model (phoneme bigram model) L_A are trained on the speech of accent *A*. During testing a Viterbi decoder finds for each recognizer the most likely state sequence representing the speech utterance incorporating the HMM and language model and assigns the log likelihood scores $S_A = \log P\langle O \mid \Lambda_A, L_A\rangle$ to the proposed phoneme sequences.

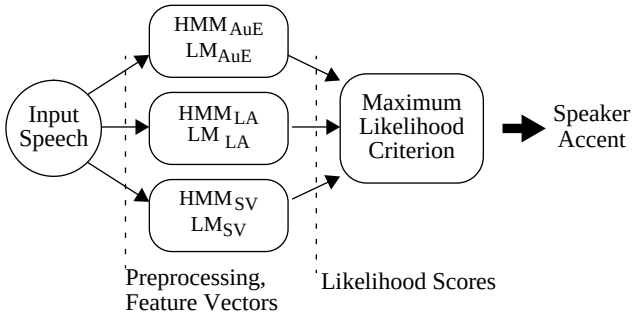

Figure 1: Accent classification system, block diagram

Prior to the final classification, a recognizer dependent bias is subtracted from the likelihood scores of each recognizer. This bias is due to the limited size of the training databases and estimated as the average of the likelihood scores for the utterances processed by the recognizer. The maximum likelihood criterion is then applied to choose the recognizer with the highest likelihood score as the most probable to represent the accent of the test utterance:

$$\hat{A} = \arg\max_{A} \{\log P\langle O \mid \Lambda_A, L_A\rangle\} \quad A \in \{\text{AuE, LA, SV}\} \quad (1)$$

Phonemic transcription of the accented training data set was required for the training of the accent-specific recognizers. It was also desirable to keep the system architecture simple and portable to perform similar accent identification tasks on other languages.

Generally, we can not expect to find large phonetically labelled databases of foreign accented speech in different languages. It is feasible to collect accented speech automatically, for instance on telephone channels, but the training of the accent classification system should not rely on manual labelling.

Accented speakers will modify the articulation of the target language to a certain degree by substitutions and approximations from the phoneme set of their first language. Although these deviations from the phoneme set of the target language would be especially useful for the identification of the speaker accent, their representation in a phonemic transcription of the accented database would be very complex and difficult to achieve.

We therefore focussed on the phoneme set of the target language (AuE) using automatic speech segmentation. A continuous speech recognition system was not available but the orthographic transcription of the ANDOSL sentences (see Section 2.2.) allowed for forced phoneme alignment with an AuE phoneme segmentation system, developed for the automatic transcription of the ANDOSL database. The phoneme segmenter uses a set of 44 AuE phoneme HMMs and two silence models that were trained with Baum-Welch reestimation on phoneme segments on manually labelled speech.

Due to the limited amount of accented data available for the training of the accented recognizers some of the HMM models were unstable or the training did not converge. We therefore obtained the accent specific HMM sets by retraining the AuE models from the phoneme segmenter with embedded Baum-Welch reestimation on accented data. Skipping the model initialization and the first iterations of Baum-Welch reestimation also reduced the computational load for training. Figure 2 summarizes the steps of the training procedure.

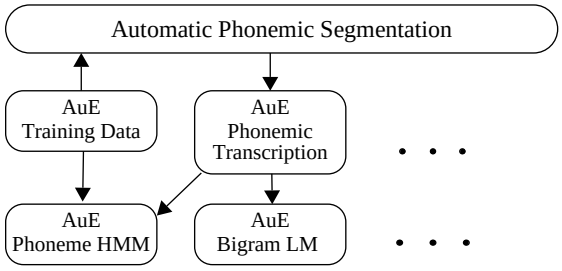

Figure 2: Training of accent-specific recognizers, block diagram

The entries of the phonemic pronunciation dictionary used in the forced speech segmentation and phoneme label alignment contained optional phoneme cluster substitutions and insertions. These captured a large amount of the pronunciation variability found in AuE speech in general and in the accented speech.

Characteristic pronunciation variations in accented speech include the change of stress patterns in the utterance, which often cause vowel substitutions. The dropping of word final consonants like /d/ and /t/ is due to more or less casual speaking style. Speech produced with high speed is often characterized by deletions of consonants, especially within consonant clusters, and the reduction of short function words. Typical for Australian English is also the reduction of unstressed vowels to the /schwa/ sound. Other variations are due to the regional differences of AuE and its dialects, which are grouped into broad, general and educated. Examples are the optional insertion of consonants in homorganic nasal-fricative clusters (produced with the same organs of the vocal tract) as well as vowel and consonant clusters. Multiple pronunciations in the dictionary entries are also provided for words of foreign origin.

Accent dependent phoneme bigram language models were trained in order to investigate whether the phonotactic variations in the continuous accented speech make use of the flexibility in the pronunciation dictionary entries in a systematic and accent dependent way.

2.2. Accented Database

The speech corpus is part of the Australian National Database of Spoken Language (ANDOSL [4]). This accent classification study has been confined to sentences read by male speakers from three accented speaker groups: native Australian English (AuE), Lebanese Arabic (LA) and South Vietnamese (SV). The accented speakers belonged to different age groups and had spent different amounts of time in Australia. Their proficiency in reading and speaking English varied but was not discriminated in this study. The amount of data available increased during the development of the

accent identification system. Table 1 shows the final corpus size and Section 3 outlines the assignment to system training and testing.

	AuE	LA	SV
No. of speakers	17	26	24
No. of utterances	1150	1450	1350
Average utterance length (sec)	3.2	5.5	5.2

Table 1: Accented database overview

Over 2000 sentences from 16 separate AuE ANDOSL speakers were used to train the automatic AuE phoneme segmenter.

2.3. Preprocessing and Modelling

The ANDOSL database had been recorded in a studio with 20 kHz sampling frequency. A 16 ms Hamming window and preemphasis ($\eta = 0.97$) was applied for feature extraction with 5 ms frame advance. The feature vectors consisted of 12 MFCC coefficients, 12 delta MFCC coefficients, log energy and delta log energy.

All HMM phoneme models have a 3 state left-to-right topology with skip transition. The models are context independent and the probability distributions for the state occupation by the feature vectors are modelled with 3 to 6 Gaussian mixtures with diagonal covariance matrices. The Entropic HTK-Toolkit V1.5 was used for signal processing, training and testing of phoneme HMMs and bigram phoneme language models.

3. EXPERIMENTAL RESULTS

This section reports the results of the foreign accent identification experiments during the stages of the system development. The system was used off-line and the accent-specific recognizers processed the test utterances in series rather than in parallel.

3.1. Database Segmentation

The automatic segmentation of the accented database limits the accuracy of the phoneme boundaries. This disadvantage is partially offset by the fact that automatically generated labels are more consistent than labels produced manually by different phonetic experts.

	AuE	LA	SV
Matching labels, (%)	91.0	82.6	65.6
Boundaries within 5 ms, (%)	38.8	33.2	29.2
Boundaries within 16 ms, (%)	82.2	80.0	72.3
Boundaries within 32 ms, (%)	94.5	91.7	87.0

Table 2: Automatic phoneme label alignment. Percentage of automatically generated phoneme labels matching manual transcription;(%) label boundary differences within 5, 16 and 32 ms.

The quality of the automatic alignment of phonemic transcription to the accented speech with the Australian English phoneme segmenter was measured on 600 sentences from 3 test speakers with AuE, LA and SV accents from the ANDOSL corpus (Table 2).

3.2. LA versus SV Discrimination

The first accent classification experiment was set up to investigate whether two accented speaker groups, whose first languages were LA and SV, can be discriminated based on the proposed maximum likelihood criterion. In a multi-speaker test 50 utterances from each of 25 LA and 23 SV male speakers were used for training and testing. The classification accuracy averaged over both accents for the test on 10 utterances from each speaker was 95.2% using only the accent-specific phoneme HMMs with 3 Gaussian mixtures in the recognizers. Incorporating the accented bigram language models in the Viterbi decoding yielded 97.4%.

A speaker independent test was performed by cycling of the training and testing procedure through the data, excluding the test speaker from training. 50 HMM and bigram model sets were trained. Again 480 utterances were tested resulting in an average accent classification rate of 78.4% and 80.1% without and with phoneme bigrams. The increased complexity of the task resulted in a reduced classification rate compared with the multi-speaker test, but the results clearly show the potential of the likelihood scores from the accent-specific recognizers for accented speaker group discrimination.

3.3. Speaker Independent Test AuE, LA, SV

The next experiment was designed to evaluate the accent identification method on the discrimination on 50 utterances from each of 15 AuE, 25 LA and 23 SV accented speakers. Training and testing again cycled through the database increasing the processing load by training 66 separate recognizers. Accent classification was performed on accent pairs as well as on all 3 accent classes. The best results in the 3 class discrimination were achieved by firstly processing the test utterance through the accent pair classifiers and then applying the maximum likelihood criterion to the output scores from these classifiers. The recognizer dependent bias was estimated separately for each of the accent pair classifiers.

Accent pairs	AuE/LA		AuE/SV		LA/SV		Average
	AuE	LA	AuE	SV	LA	SV	
HMM	76.3	80.1	88.1	87.9	81.4	84.6	83.1
HMM + LM	80.9	82.2	92.0	90.1	81.5	85.0	85.3
3 classes	AuE		LA		SV		Average
HMM	76.7		65.9		78.9		73.8
HMM + LM	82.7		67.4		79.8		76.6

Table 3: Accent discrimination for AuE, LA, SV; Accent classification rate (% correct) using accent-specific phoneme HMMs with and without phoneme bigram language models.

Table 3 summarizes the results. The accent pair classifiers processed between 2000 and 2400 utterances each. Averaged over all accent pairs the speaker accent was recognized correctly for 83.1% and 85.3% of the utterances without and with bigram language models. The classification between AuE and LA speakers appeared to be more difficult than the discrimination within the other accent pairs. The accent classifier for all 3 speaker groups was

tested on 3150 utterances and reached 73.8% and 76.6% average accent classification rate without and with bigram language models.

The results show that the incorporation of the accent specific phoneme bigram language models contributes only marginally to the performance of the accent classification system. There are several reasons. Firstly, the pronunciation dictionary of the automatic phoneme labeller does not provide enough flexibility to cover all the variations in pronunciation that are systematically different between the accented speaker groups. Secondly, automatic segmentation using the AuE phoneme set inadequately identifies phonemes which accented speakers import from their first languages. Finally the amount of training data be may too limited to train stable phoneme bigrams. There is also a strong bias in the bigram estimation due to the fixed content of the read sentences set in the ANDOSL database.

3.4. Classifier Training with Automatically versus Manually Labelled Data

In the final stage of this study a substantial part of the accented database was manually transcribed with phoneme labels allowing for the comparison of accent classification systems trained from automatically and manually segmented data.

As the cycling through the data was too computationally expensive and not directly applicable to a front-end architecture, the database was split into speaker independent training and test sets (Table 4). This reduced the amount of training data for the accent classification system compared to Section 3.3.

(a) Manually segmented data	AuE	LA	SV
No. of speakers; Training/Testing	7/15	6/20	9/15
No. of utterances; Training/Testing	650/2700	450/600	600/450
(b) Automatically segmented data	AuE	LA	SV
No. of speakers; Training/Testing	10/5	15/10	15/8
No. of utterances; Training/Testing	2000/900	750/300	750/240

Table 4: Speaker independent assignment of accented data to training and test sets for the accent classification system with (a) manual and (b) automatic phoneme segmentation.

The phoneme label distributions were highly correlated between the automatic and manual transcription of the data (AuE: ρ=0.972, LA: ρ=0.872, SV: ρ=0.835). The average phoneme durations were also highly correlated (AuE: ρ=0.994, LA: ρ=0.997, SV: ρ=0.998). The automatic segmentation inserted additional short phoneme and silence labels, thereby reducing the average phoneme duration.

	manual segmentation		automatic segmentation	
	accent pair	3 accents	accent pair	3 accents
HMM	83.7	73.1	82.8	71.1
HMM + LM	84.2	73.2	83.2	71.8

Table 5: Classifiers trained from manually vs. automatically segmented data; Average accent classification rate (% correct).

Table 5 shows the results for the accent discrimination with the two accent identification systems. The reduced amount of training data resulted in a performance reduction with respect to Section 3.3. The best results for the 3 class discrimination were achieved when applying the maximum likelihood criterion directly to the likelihood scores from each of the accent-specific recognizers.

4. DISCUSSION

We have presented an approach to the development of a foreign accent classification for continuous speech based on phoneme segmentation. The results have shown that the likelihood scores produced by accent-specific phoneme recognizers can be used to discriminate between speaker accents. The accent-specific phoneme bigram language models did not contribute as significantly as expected, partly due to the limited training data set. Also the performance of the accent classification is very sensitive to the amount of and speaker variety captured in the training data. The comparison of accent classifiers trained from automatically and manually segmented data outlined the effectiveness of the automatic segmentation with respect to the classification task.

In future work we will expand the classification algorithm beyond the maximum likelihood criterion used here. We will focus on more discriminative classification techniques on a phoneme by phoneme basis. We also plan to establish some benchmarks for speaker accent classification based on a human perception study.

5. AKNOWLEDGEMENTS

We thank Julie Vonwiller, Chris Cleirigh, Inge Rogers and Wendy Lewis of the Speech Technology Research Group for preparing the manual transcription for a large part of the database. We are grateful to Marc Zissman for useful suggestions for the experimental design.

6. REFERENCES

1. Blackburn, C. S., Vonwiller, J. P., King, R. W., “Automatic Accent Classification Using Artificial Neural Networks”, Proc. of Eurospeech ‘93, Vol. 2 pp. 1241-1244.

2. Hansen, J. H. L., Arslan, L. M., “Foreign Accent Classification Using Source Generator Based Prosodic Features”, Proc. ICASSP ‘95, pp. 836-839.

3. Itahashi, S., Yamashita, T., “A Discrimination Method Between Japanese Dialects”, Proc. ICSLP ‘92, pp. 1015-1018.

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

5. Zissman, M. A., Singer, E., “Automatic Language Identification of Telephone Speech Messages Using Phoneme Recognition and N-gram Modelling”, Proc. ICASSP ‘94, pp. 305-308.

6. Zissman, M. A., “Language Identification Using Phoneme Recognition and Phonotactic Language Modelling”, Proc. ICASSP ‘95, pp. 3503-3506.