A COMPARISON OF L1 AND AFRICAN-MOTHER-TONGUE ACOUSTIC MODELS FOR SOUTH AFRICAN ENGLISH SPEECH RECOGNITION

Janus D Brink, Elizabeth C Botha

Department of Electrical, Electronic and Computer Engineering
University of Pretoria, Pretoria 0002, South Africa.

ABSTRACT

Speaker accent influences the performance of automatic speech recognition (ASR) systems. Knowledge of accent based acoustic variations can therefore be used in the development of more robust systems. The goal of this project is to characterize the vowels and diphthongs of second language (L2) South African English to aid in the adaptation of existing first language (L1) English recognition systems for better L2 performance. This paper investigates the differences between the vowels and diphthongs of L1 and L2 English in South Africa and is specifically aimed at L2 English speakers with a native African mother tongue for instance isi-Xhosa, isi-Zulu, Tswana or Sepedi. The vowel systems of English, and African languages, as described in the linguistic literature, are compared to predict the expected deviations of L2 South African English from the L1 norm. Acoustic models based on formant and Mel-scaled cepstral features of 80 context dependent phonemes from L1 and L2 speakers are compared. Our findings agree well with those linguistically predicted, in particular, evidence of equivalence-classification, peripheralization of schwa and changes in diphthong strength are observed.

1. INTRODUCTION

Dictation and other automatic speech recognition (ASR) systems are trained for a specific language dialect. The performance of such ASR systems is therefore sensitive to accents of the language it was trained on. In South Africa, where numerous cultural and ethnic groups are present, the accents of second language (L2) English vary significantly from the first language (L1) norm. A study was therefore performed to quantitatively determine some of the acoustic differences between the vowels and diphthongs of L1 and L2 South African English (SAE) in formant space as well as the Mel-scaled cepstral domain [1]. The second language speakers of interest were native speakers of African languages, for instance mother tongue speakers of isi-Zulu, isi-Xhosa or one of the Sotho languages. The goal of this study is to form a statistical model of the acoustic differences in pronunciation of several speech sounds in L2 English as compared to their L1 counterparts in order to improve the performance of ASR systems in South Africa.

Section 2 briefly examines some linguistic background on the English vowel system as compared to those of the African languages and predicts African L2 English linguistically. Section 3 describes the data set, while Section 4 details the methods used to analyse the data. Results of the analysis are discussed in Section 5.

2. LINGUISTIC ACCENT

The vowel system of English [2, 3] can be linguistically compared to that of native South African languages [4, 5] in order to predict mother tongue influence on L2 English learning. In English the stressed vowel system includes at least fourteen commonly used vowels, excluding the rounded variants and diphthongs. The Nguni languages (including isi-Zulu, isi-Xhosa and Swazi) use five main vowels, while the Sotho languages (including Sepedi, Ndebele and Tswana) use seven main vowels, all of which are peripheral in the absence of central /a/.

Previous studies on African accents of English [6, 7, 8, 9] have shown that central English vowels (/a/, /e/, /o/ and /ɔ/ as in but, bird, a and nurse) are avoided and tend to more peripheral vowels (/ɛ/, /ɛ/, /ɔ/, /ɔ/, /ɔ/).

The equivalence classification theory proposed by Flege [10] is also responsible for drastic differences between L1 and L2 pronunciation, where unfamiliar L1 English phonemes are substituted by the closest mother-tongue phoneme. This affects the L2 pronunciation of vowels like /æ/, /ʌ/, /ɔ/ and /ɔ/ which are not valid phonemes in the native Southern African languages [11, 5, 12].

Another tendency is for diphthongs to be monophthongized, especially shorter, closing diphthongs (/ɒ/, /ʌ/) where the second element is hardly heard [13, 9]. Longer diphthongs are preserved, but both elements receive the same emphasis, where in L1 English emphasis on the second element is diminished. The mentioned decentralization of vowels affects all centering diphthongs (/ɑ/, /ɛ/, /ɔ/) and tends to make them opening diphthongs (/ɑ/, /ɛ/, /ɔ/).

3. DATA

A database of first and second language South African English was compiled from broadcast speech, obtained from local television news and programmes. All data was recorded onto VHS video tape, where after it was digitized at 22.050 kHz with 16-bit precision. Continuous speech phrases of more than 38,000 words from 236 speakers were recorded, digitized and orthographically transcribed. Each speaker was subjectively scored by the first author on a 5-point scale with L1 on the one end and L2 on the other; borderline speakers were ignored.

An intersection between the two language groups was extracted to include only those words for which a representative sample set was present in each - i.e. utterances from at least 10 different speakers per language group. This intersection was considered separately for each gender, since comparing within gender groups favorably decreases variance of the data within a mother-tongue
4. EXPERIMENTAL PROCEDURE

The data was analysed in two measurement spaces. Firstly, a formant-space comparison was made, which can be correlated directly with the linguistic predictions. This was followed by a hidden Markov model based analysis in the Mel-scaled cepstral domain [15], which is one of the popular methods used by modern automatic speech recognition systems. The subsequent sections detail the analysis methods used in each domain.

4.1. Formant analysis

The Split Levinson algorithm [16] was used to estimate the first three formant values of the voiced speech segments over time, forming formant tracks. The static vowels were modelled as normal distributions, where the mean and variance for each vowel were calculated. Analysis of variance (ANOVA) [17] was used to compare the L1 models with their L2 counterparts to test if they differed significantly (with respect to the 95th and 99th percentile levels) from one another. If the distributions differed significantly in any of the dimensions, the groups were considered statistically different. The movement of the vowel means was examined in the F1-F2 formant plane and compared to the linguistic predictions.

Diphthongs were modeled by 3-state, single-mixture, left-to-right hidden Markov models (HMMs). This allows the HMM to model the diphthong at three stages. The onset, the center and the closure of the diphthong are modeled as separate Gaussians (similar to the three-section spline model in the formant space). The diphthongs were analysed in two ways. Firstly L1 diphthongs were compared to their L2 counterparts by simply comparing the corresponding three Gaussians, again using the ANOVA technique.

Secondly, diphthongs were analysed for the effects of diphthongization and monophthongization. This means the relative strength of a diphthong needs to be measured, where the “strength” indicates the amount of change the vocoid undergoes over time as the diphthong is pronounced. Such a relative measure was obtained by comparing the three HMM-state Gaussians of a particular language group’s diphthong model with one another, treating them as a three-class ANOVA problem. This gave an indication of the differences between the three stages of the diphthong as modelled by the HMM. When the L2 diphthong has a higher strength value than its L1 counterpart, it indicates diphthongization, while a lower value indicates monophthongization.

5. RESULTS AND DISCUSSION

The analysis of variance results of the vowel comparisons are given in Tables 2 and 3. The first column indicates the ARPBET symbol of the vowel, while the second column indicates the word context. This is followed by the degrees of freedom of the ANOVA test (DOF = total number of utterances in both language groups - 2). The next three columns indicate the ANOVA F-ratio in each of the formant dimensions: a light-grey block indicates that the two distributions differed significantly to a 95% level of certainty, while a dark-grey block denotes a significant difference to a certainty of 99%. The formant result columns are concluded by the HMM. When the L2 diphthong has a higher strength value than its L1 counterpart, it indicates diphthongization, while a lower value indicates monophthongization.

4.2. HMM analysis

In the Mel-scaled cepstral domain, a thirteen dimensional feature vector was computed for each 10ms window, with a frame-advance of 5ms (50% overlap) over the duration of the utterance. This resulted in a $13 \times n$ matrix, where $n$ represents the number of frames available from the sample. Each vowel was then modeled by a very simple single-state, single-mixture, left-to-right Markov model, resulting in the estimation of a single 13-dimensional normal distribution. Each L1 vowel model was compared to its L2 counterpart by performing analysis of variance on the state variables of the Markov models, resulting in a significance score for each of the 13 dimensions.

Diphthongs were modeled by 3-state, single-mixture, left-to-right hidden Markov models (HMMs). This allows the HMM to model the diphthong at three stages. The onset, the center and the closure of the diphthong are modeled as separate Gaussians (similar to the three-section spline model in the formant space). The diphthongs were analysed in two ways. Firstly L1 diphthongs were compared to their L2 counterparts by simply comparing the corresponding three Gaussians, again using the ANOVA technique.

Secondly, diphthongs were analysed for the effects of diphthongization and monophthongization. This means the relative strength of a diphthong needs to be measured, where the “strength” indicates the amount of change the vocoid undergoes over time as the diphthong is pronounced. Such a relative measure was obtained by comparing the three HMM-state Gaussians of a particular language group’s diphthong model with one another, treating them as a three-class ANOVA problem. This gave an indication of the differences between the three stages of the diphthong as modelled by the HMM. When the L2 diphthong has a higher strength value than its L1 counterpart, it indicates diphthongization, while a lower value indicates monophthongization.
Analysis of variance results for the vowels /æ/ (/AA/) through /æ/ (/AXR/).

99% level of certainty in four of the cepstral dimensions, resulting in a difference score of 8. Other strongly affected vowels include /œ/ /œ/ /æ/ /æ/ and /œ/ (/OW/, /UW/, /IH/, /AX/ and /AXR/), where large differences in F2 are also seen that translate to significant differences in many of the cepstral dimensions. The vowels /œ/ /œ/ and /œ/ (/AO/ and /IY/) seem to be affected differently, where the long and short forms of certain vowels like /æ/ (/EH/ and /AE/) also show significant differences in most contexts. The long and short forms of certain vowels like /æ/ and /æ/ (/AO/ and /IY/) seem to be affected differently, where the long form (/æ/ in all or /æ/ in be) is more affected by L2 accent than the short form (/æ/ in not or /æ/ in the word in). Vowels that show little difference between L1 and L2 pronunciation are the short form of /æ/ and /æ/ and most contexts of the vowel /æ/ (/IU/ and /AE/).

The diphthongs are summarized in Table 4. As in the vowel results tables, the first three columns indicate the phoneme, word context and degrees of freedom. This is followed by a set of three columns repeated for each of the three formant dimensions. Each set indicates the ANOVA results for the three spline sections of that formant dimension (light- and dark-grey blocks indicate a significant difference to a certainty of 95% and 99% respectively). The next column gives an aggregate formant score ranging between 0 and 18, calculated as before, indicating the degree of difference between two formant models. These formant results are followed by the cepstral ANOVA results. Three columns indicate the L1/L2 difference scores (ranging from 0 to 26) of the three successive HMM states. The score is calculated over the 13 cepstral dimensions for each state. Two arbitrary levels were chosen to aid in visualizing these results: a score below 8 is indicated in a white block (small difference), from 9 to 14 as light-grey (medium difference) and 15 to 26 as dark-grey (large difference). This is followed by the sum of these three scores, resulting in a relative difference) and 15 to 26 as dark-grey (large difference). This is analyzed for the closing diphthongs /a/ /æ/ and /æ/ (/IH/), which are shown as light- and dark-grey blocks respectively.

Table 2. Analysis of variance results for the vowels /æ/ (/AA/) through /æ/ (/AXR/).

Table 3. Analysis of variance results for the vowels /œ/ (/EH/) through /œ/ (/UW/).

Successive HMM states. The score is calculated over the 13 cepstral dimensions for each state. Two arbitrary levels were chosen to aid in visualizing these results: a score below 8 is indicated in a white block (small difference), from 9 to 14 as light-grey (medium difference) and 15 to 26 as dark-grey (large difference). This is followed by the sum of these three scores, resulting in a relative measure of the difference between the two diphthong models. The final three columns indicate changes in diphthong strength, with the first two columns indicating the diphthong strength measure (light- and dark-grey blocks respectively). Most diphthongs are significantly affected in L2 SAE. The largest differences are present for the phoneme /œ/ (/OW/), which was treated as a diphthong in this experiment. Small diphthong strength scores in the contexts of /œ/ /œ/ /æ/ and /æ/ (/AU/ /IA/ /IH/ /AX/ and /AXR/) indicate vowel-like pronunciation in both language groups.

The strong difference scores for the closing diphthongs /œ/ and /œ/ (/AU/ /IA/ /IH/ /AX/ and /AXR/) are due to monophthongization by L2 speakers, while diphthongization is seen for the centering diphthongs /œ/ /œ/ (/EH-AXR/)
intheir andh/ in will. In other cases, such as /aw/ (/AWR/) in
our, /aU/ (/AY/) in by and /ow/ (/OW/) in so, diphthong strength is
similar for the two language groups, but the L2 diphthong moves
in a different direction or is translated with respect to its L1 coun-
terpart. The opening diphthongs /s/aw, /s/ae/, /s/æ/ and /s/I (/WA/, /
WAE/, /WEH/ and /WIY/) are less affected.

### 6. CONCLUSION

Our results indicate that the differences in the linguistic vowel-
spaces of English and the African languages directly influence the
pronunciation of second language South African English. The
results also agree well with the linguistic characterization of L2
SAE. Significant differences between the L1 and L2 vowel and
diphthong models were found in most instances. The vowels most
seriously affected are those which do not exist in the native African
languages, where the closest vowel in the mother tongue is substi-
tuted. This also influences diphthongs, where an unfamiliar diph-
thong element is either replaced by the nearest native vowel, or
dropped. In the former case over-articulation of diphthongs can
appear, while in the latter the diphthong is reduced to a vowel.
Central vowels and centering diphthongs are avoided and tend
to more peripheral pronunciation. The behaviour of formant and
Mel-scaled cepstral models are closely related for both vowels and
diphthongs and the derived diphthong strength metric is a good
indicator of a diphthong’s change over time.

The observed L2 SAE pronunciation will certainly affect Mel-4
scaled cepstral based speech recognition systems and ways to
adapt such systems would include adapting their phoneme mod-
els where significant differences are found and adding alternative
pronunciations for these words in the lexicons.

### 7. REFERENCES

vowels and diphthongs of first and second language South
African English,” in Proceedings of the 12th Annual Sympo-
sium of the Pattern Recognition Association of South Africa,

Society of Japan, Tokyo, 1958.


[4] C.M. Doke, The Southern Bantu Languages, Oxford Uni-


[6] D. Zier, Handbook of the speech and sound changes of the
Bantu languages of South Africa, University of South


[8] J.C. Wells, Accents of English, Beyond the British Isles,


a foreign language: evidence for the effect of equivalence
classification,” Journal of Phonetics, vol. 15, pp. 47–65,
1987.

and Co., Cape Town, 1941.

ception of the vowels of Tswana-English,” South African

[13] W.L. Lanham, The way we speak, Van Schaik, Pretoria,
1967.

42, pp. 175–184, 1952.

Recognition, Prentice Hall, Englewood Cliffs, New Jersey
07632, 1993.

[16] L.F. Willems, “Robust formant analysis,” Institute for Per-
ception Research Report 529, Eindhoven, The Netherlands,

York, 1980.