Automatic systems for assistance in improving pronunciations

Jacob A.C. Badenhorst, Daniel R. van Niekerk, Etienne Barnard

Human Language Technologies Research Group
University of Pretoria / Meraka Institute, Pretoria, South Africa
s22096745@tuks.co.za, s22088840@tuks.co.za, ebarnard@csir.co.za

Abstract

Improving the pronunciations of non-native language learners is an important task in a multilingual society. We focus on segmental aspects of pronunciation, and investigate the design of automated assistants that can be used to improve (1) the articulation of phones and (2) the production of tone. Initial experiments investigated phone productions in English by speakers of Nguni languages, and the tone of English speakers when producing speech in an Nguni language. Our initial results are promising, and point to the improvements that are required to develop a practical system.

1. Introduction

Language learning is a typically contradictory human skill. Whereas the vast majority of children are able to learn a language faultlessly without specialized supervision, people find this ability to be increasingly challenging as they mature (and even the most sophisticated computer algorithms are utterly incompetent in this regard). Factors such as globalization and the social integration of societies have raised the importance of this paradox, since adults are increasingly required to learn new languages.

In this paper, we investigate one aspect of language learning, namely the pronunciation of segmental units. The inventories of such segments differ widely between different languages - for example, the Nguni languages contain "click" phonemes that are absent in the Germanic languages, whereas the Germanic languages tend to employ a wider range of vowel sounds than the Nguni languages. Another important inter-lingual difference (at the segmental level) is how factors such as tone and stress are realized - in so-called tonal languages, for example, segmental tones are used to distinguish between different words, whereas such a distinction is not utilized in stress languages. (These segmental phenomena are, in turn, distinct from supra-segmental or prosodic aspects of speech, which determine the "melody" of speech at longer time scales - words, phrases, etc.) In particular, we investigate how techniques from automatic speech recognition can be employed to assist in the improvement of pronunciations at the segmental level.

Although a significant amount of work has been done on automated tools for pronunciation learning (see [1] for an example of recent work), little is known about the specific challenges that occur in a South African context. South Africa has a unique mixture of official languages, consisting of nine languages from the Bantu family of languages and two Germanic languages. Given the significant linguistic differences between these two families of languages, it is not surprising that language learners whose early linguistic experiences are limited to one group have significant difficulties in acquiring the pronunciations of the other group.[1]

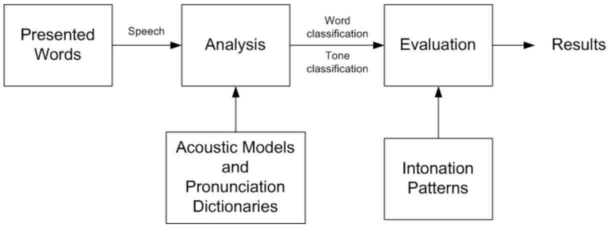

Figure 1: *General system for evaluation of speech segments*

For our initial work, we focus on English as a representative of the Germanic languages, and Nguni languages (isiZulu and isiXhosa) as representatives of the Bantu family of languages. Specifically, we investigate the articulation of English phones by speakers of isiXhosa, and the production of isiZulu tones by English and Afrikaans speakers, and present systems that can be used to evaluate pronunciations in these two contexts. In Section 2.1 we outline the general approach that was taken for these two tasks, and Section 3 describes the details of systems that were developed to implement this approach. Section 4 contains initial results obtained with these systems, and Section 5 summarizes our main conclusions.

The contributions of this research are twofold: on the one hand, we show a basic system design that can be used to detect pronunciation errors of the type that are expected to occur in the South African context. On the other hand, we provide initial indications of the accuracies that can be achieved using this design (which in turn suggests the areas most in need of further refinement).

2. Approach to the evaluation of segmental pronunciations

2.1. Approach

A general system that can be used to evaluate the pronunciation of speech segments is shown in Figure 1.

Words presented to the system are obtained from predefined word lists. The correct speech data for the presented word is found and the analysis shown in Figure 2 is performed. Different acoustic models and pronunciation dictionaries are used as required by this step. The appropriate tone and word classifications generated are evaluated to obtain suitable results. For tone

[1] Whether it is, in fact, desirable for language learners to acquire such cross-family pronunciation skills has become a politically charged topic in certain circles. For our purposes, it is sufficient to note that a certain degree of cross-family pronunciation acquisition is important for basic intelligibility, and is therefore uncontroversially useful.

classification further evaluation using intonation patterns takes place.

We now discuss some of the details of the phone-articulation and tone-production subsystems.

2.2. The articulation of phones

Pronunciation error detection is concerned with determining if a speaker has uttered a phone differently to the pronunciation that was expected given a specific context. In our system design, known words (the context) will be presented to the user and the system should measure how well the received input matches what was expected.

The most straightforward approach to this task would be to use a general-purpose speech-recognition system with a large vocabulary and unrestricted grammar to recognize the spoken utterance. However, general speech recognition is currently not sufficiently accurate to be used in this manner. Forced alignment is therefore typically employed in systems of this nature [2]. That is, the input utterance is assumed to correspond with the requested words presented to the speaker, and the speech-recognition system finds the best mapping (in time) between the acoustics produced by the speaker and these words.

2.2.1. Identifying most problematic phones

The first technique that addresses the problem of analysis is to target only those phones that are most significant for a particular learning task[1]. To select those phones, it is important to consider the characteristics of the specific group of speakers to which the system is targeted. As discussed above, we focus on the pronunciation of English by first-language speakers of isiXhosa. We have therefore analyzed a set of recordings made by first-language isiXhosa speakers reading English texts.[2]

An inspection of the above mentioned recordings clearly indicated the most problematic phones. As expected from the differences in the phonetic inventories of these two languages, speakers had most difficulty in producing the English vowels. An inspection of the above Our subsequent attention therefore focused on the articulation of vowels.

2.2.2. Identifying mispronounced vowels

Two approaches to the identification of mispronounced vowels were investigated.

- One approach took the acoustic scores, obtained by forced alignment of the correct pronunciations to the input utterance, as indication of the quality of vowel pronunciation. These scores are the average log likelihoods computed by Hidden Markov models trained on data from first-language English speakers, and are expected to correlate well with the accuracy of the pronunciation
- An alternative approach is to include the expected mispronunciations in an alternative pronunciation dictionary. This specialized new dictionary is used in conjunction with the normal pronunciation dictionary, and whenever the erroneous pronunciation is preferred by forced alignment, a possible pronunciation error is indicated.

For the second approach, the generation of appropriate variations on the pronunciations of words is of great importance. One way to achieve this is to use phonetically transcribed data of second language speakers. The transcriptions will typically employ an inventory that includes all the phones found in the data (which may be more than the actual phones of the target language). However, by mapping these phones to the set of phones found in the target language, it is possible to produce pronunciation alternatives using only models of the target language.

2.3. Tone production

In tone languages, lexical tone can be used to attach different meanings to words which consist of the same sequence of phonemes. Since this use of tone may not be familiar to first-language speakers of non-tonal languages, it is expected that such speakers may have problems in producing appropriate lexical tones.

The main components determining prosody are intonation (referring to the variation of voice pitch) and stress (referring to the intensity with which certain syllables are realized). These two subjective elements correspond well to the measurable parameters, namely fundamental frequency (*f0*) and signal intensity (or energy) respectively. These same two factors play a major role in determining lexical tone, and are used in our analysis of tone production. Since tone is associated with specific syllables in an utterance, the first problem that needs to be addressed is that of segmenting the data into its syllabic constituents. For this task, we use forced alignment (as described in Section 2.2) to first find phone boundaries. The phones are then grouped together (using a simple set of rules developed for an isiZulu TTS system [3]) to form syllables. Finally, a tone classification algorithm is employed to determine whether appropriate tone levels had been produced by the speaker.

There are two major factors which will determine how accurately such an evaluation of tone production functions, namely *segmentation accuracy* and *the accuracy of tone classification*. To evaluate segmentation accuracy, we compare the output of the system for a given reference set to data which have been aligned manually. Various methods exist for this comparison, including calculating the mean square error of the boundary locations or counting errors based on a predefined threshold. However, these methods fail to take into account the seriousness of errors based on the relative distances between boundaries. A confidence measure proposed by Paulo and Oliveira [4], considers segmentation results by determining an "Overlap Rate" for each phone in the reference data set, which can be expressed as a percentage and thus gives an objective measure of alignment performance, which takes into account differing phone lengths. We employ this measure of segmentation accuracy.

Tone classification is similarly assessed by comparing manually assigned tones to those that are computed automatically. Here, there are two issues.

- The appropriate set of labels to use for a given language is sometimes not entirely clear. For the Nguni languages, for example, it is clear that high (H) and low (L) tones should be distinguished, but the linguistic status of falling (F) tones is the topic of some debate. [5]
- Even when a labeling convention is determined, there is a fair degree of subjectivity in assigning tones to physically realized utterances. Factors such as dialect, pragmatics, emotion and speaker idiosyncracies can all influence the tones produced by a speaker, and the same factors will therefore influence any transcriptions used for comparative purposes.

For our work, we have chosen a particular tone-labeling con-

[2] These recordings and their transcriptions were provided by Dr. Febe de Wet of the University of Stellenbosch.

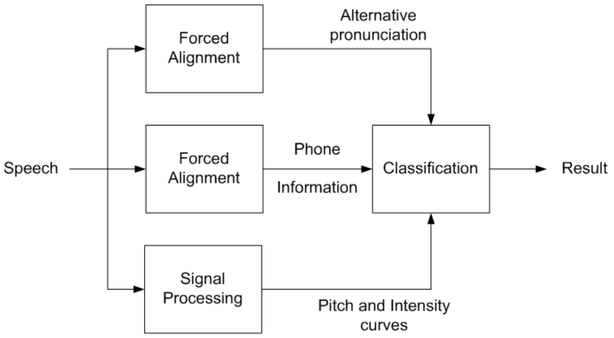

Figure 2: *Pronunciation and Intonation analysis*

vention (see below), and accept that subjectivity will limit the accuracy of tone classification that can be achieved.

3. Experimental systems for pronunciation assistance

Figure 2 shows the main components of a system that can be used for pronunciation evaluation. Below, we describe how the various subsystems are implemented, using the open-source software package HTK [6].

3.1. Acoustic score

When applying forced alignment, using HTK's *HVite* tool, a common way of determining how closely the input correlates with the transcription provided, according to the trained models, is by way of the acoustic score. The acoustic score represents the sum of the log likelihoods for all the frames in a specific speech segment [7]. This acoustic score value represents the likelihood that a speech segment represents a particular symbol according to the statistical models (HMMs). The acoustic score however, cannot be directly used as a means of comparing speech segments, as the value depends on the length of the segment. HTK does provide a means of normalizing this acoustic score for a particular segment, which simply entails dividing the score by the number of frames contained in the segment. This value thus represents the average log likelihood per frame for the given segment, and can be used to compare speech segments of different lengths (typically different phones) to determine which segments fit better.

3.2. Selecting appropriate training data

Due to the small amount of training and test data available to us, it was necessary to perform tests in order to select an appropriate data set. We limited our evaluation to a number of straightforward comparisons.

The two assistants proposed in this work are intended for first-language isiXhosa speakers who are speaking English and first-language English speakers who are speaking isiZulu, respectively. That determines the test data that was used in our evaluations. For training the acoustic models, however, greater flexibility exists: since those models are only used for forced alignment, it could be that a large corpus of speech in a different accent group could be preferable to limited corpora in the two accent groups in our focus (for which only limited data is available to us).

The following data sets were therefore investigated for HMM training:

1. *TIMIT* - The HMMs were trained with speech data from the DARPA TIMIT speech database, using the Carnegie Mellon University (CMU) pronunciation dictionary. A rudimentary phone set mapping was applied in order to represent all isiZulu words with the CMU phone set. In comparison with our other corpora, this is a very large set of utterances, but the dialect of English not strictly applicable to either of our tasks.
2. *ZULU* - The HMMs were trained with isiZulu speech data from a group of ten male isiZulu speakers, using the *Buhle* phoneset used in the isiZulu TTS system at the *Meraka Institute*.
3. *HYBRID* - The HMMs were trained with a combination of the TIMIT corpus and the limited isiZulu corpus, using a phone mapping to CMU phones where possible, but retaining unique phones from the *Buhle* phoneset.

A simple test set consisting of a few isiZulu words containing a reasonable phonetic diversity and pronounced by a native male isiZulu speaker was employed to determine the performance of the HMM systems. This was done by computing the average log likelihood per phone for each word and comparing these values. The same test set was also used to determine the phone boundary accuracy resulting from the different training sets, using the overlap measure describe in section 2.3. These results are summarized in Table 1.

Data set	Avg. log likelihood	Overlap rate
TIMIT	-94.48	51.96
ZULU	-80.43	53.37
HYBRID	-83.34	30.32

Table 1: *Average likelihoods and overlaps of words in an isiZulu test set, using different training data sets*

These two experiments suggest that the *ZULU* data set yields the best results in terms of both identifying the individual phonemes and boundary determination accuracy. In the case of the relevance of the phonemes, the *ZULU* data set clearly fits better with the reference isiZulu utterances and thus using this system as a measure of the input relevance is justified. When one considers the boundary accuracy results, the average “Overlap Rate” for the *TIMIT* data set is close to that of the *ZULU* data set. The results from the *ZULU* data set are, however, more consistent, as the boundary accuracy does not vary as much as that of the *TIMIT* data set. A possible cause for this behaviour is that the phone mapping to the CMU phoneset (used by the *TIMIT* data set) works well for certain common phonemes and thus the system benefits in this case from the greater amount of acoustic training data, but in the cases where the phone mappings are not ideal, this impacts the performance significantly. When considering the *HYBRID* data set, it seems that training acoustic models of phones that overlap when both English and isiZulu speech data is used, tends to improve the overall acoustic fit of the models. Unfortunately this results in less accurate boundary detection.

3.3. Phone pronunciation

A pronunciation test system was constructed that utilizes forced alignment on two pronunciation dictionaries. One of the dictionaries consisted only of correct pronunciations, while the other consisted of the correct pronunciation and some extra variations (alternative pronunciations). Depending on the techniques involved, one or both of the forced alignment results were used. Preliminary decisions about the pronunciation correctness is obtained form the alignment with only the correct pronunciation dictionary. Further refinements of the preliminary results are carried out when the second alternative dictionary is taken into account as well.

3.3.1. Evaluation

The test system used for pronunciation error detection utilizes the average log likelihood output values of the forced alignment to evaluate how well the expected phones match the received acoustic evidence. Expected phones are obtained from previously constructed pronunciation dictionaries. For evaluation a simple threshold value is used to measure the phone score. When a particular phone score exceeds this predefined threshold value, the phone is rejected and classified as a wrong pronunciation. There is a problem with the pronunciation of a word when one of its phones is in error. For the technique that targets only problem phones, however, a word is classified as wrong only when the pronunciations of one or more of the problem phones are incorrect.

All results were generated on a word level. It is advantageous to concentrate on pronunciation errors at the phone or word level rather than at the sentence level, because of the reliability and validity requirements. This can be achieved by targeting sub-units such as particular phoneme sets rather than judging an entire sentence as an amalgam [8]. The actual pronunciation of the word that the second language speaker uttered is found in the transcribed data. For comparison and measurement all of the data words are classified beforehand, using a similar scheme as during the generation of the system results. The output of the system is then compared with the classified data and system performance parameters are subsequently calculated.

3.3.2. Acoustic Models for Articulation Evaluation

It is important to note that the HMMs of the test system were trained using the DARPA TIMIT speech database. This corpus uses an American English phone system. However the test data of our second language English speakers is South African English, which is much closer to the British English phone system. To compensate for the differences between these phone systems, the British English pronunciations were mapped on a phone-by-phone basis to the closest phones found in the American system. This leads to different combinations of the American phones that make up closely related British pronunciations of the words used by the system. (Improved performance is expected if South African English data is used for training the acoustic models; such training is currently being undertaken.)

3.4. Automatic tone classification

An automatic tone classification system was implemented for the isiZulu language, based on research into computational models for prosody in the Nguni languages [9]. Such a system demonstrates the possibility of automating statistical classification (which is a tedious task by hand) and presents a general framework which can be re-used to investigate computational tone, stress and prosodic models for other languages.

The system utilizes three complementary components, namely phone boundary detection, pitch extraction (or fundamental frequency estimation) and statistical analysis.

Phone boundary detection is achieved with forced alignment, as described in section 2.3. The acoustic models used for alignment were trained and applied (using HTK) from a small corpus of speech data consisting of 10 native male isiZulu speakers, collected and orthographically transcribed as part of this study (see Section 3.2). For comparative purposes, we have also evaluated our system using the TIMIT-trained acoustic models mentioned in Section 3.3.2.

In order to validate the input data, it is subjected to a filtering process whereby the average log likelihood per phone is examined. This value is required to exceed a certain experimentally determined threshold value. In addition to this, the average log likelihood on a per word basis is also calculated and compared with a similar threshold to ensure that each word complies with a certain minimum acoustic fit parameter. It follows that the input signal integrity is ensured to a certain degree. These thresholds were determined by examining the average scores achieved when a test set of native isiZulu speech data was subjected to a forced alignment with reliable transcriptions (Refer to section 3.2).

Based on experiments done by Govender [10] for the Nguni languages, the algorithm proposed by Boersma [5] and source code from Praat [11] were selected to implement a simple tool to extract pitch and intensity contours from a given speech waveform.

Intonation classification can be done in many different ways, based on the particular language in question and the relevant parameters. However, we have developed a framework that can function in a similar manner for many different language applications. The basic idea is to parse sentences or phrases into words and words into syllables (if one assumes that tone and prosodic patterns are reflected on a per syllable basis), using the output of the phone boundary detection system. Statistical information from the observed pitch and intensity contours can then be generated on sentence/phrase, word and syllable levels. In the system implemented for isiZulu, such a general framework was written and a classification algorithm proposed by Govender [10] was implemented. The classification was achieved by comparing the relative differences between average pitch and intensity values to threshold values.

4. System measurements and experiments

4.1. Detecting pronunciation errors

As with most detection problems, the detection of pronunciation errors involves a trade-off between two types of misrecognitions: *false accepts* are mispronunciations that are not identified by the system, and *false rejects* are correctly pronounced words which are erroneously flagged as mispronunciations. By adjusting a detection threshold, these two types of errors can be traded off against one another.

The results in Figure 3 allow us to infer both of these error rates, when using only the acoustic scores to accept or reject words. The graphs show the percentage of words that are flagged as mispronounced through a range of threshold values (on the negative of the log likelihood). We see that all words are rejected for thresholds less than about 80; thereafter, erroneous words (as determined from the transcriptions) are slightly more likely to be rejected, but this difference is small.

Some improvement is obtained by only rejecting words based on known problematic phones, as discussed above. As shown in Figure 4, the system is about 10 % more likely to reject erroneous rather than correctly spoken words, for a range

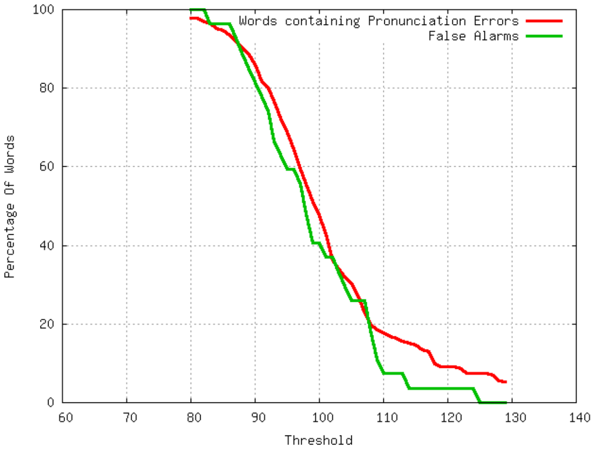

Figure 3: *Detection of erroneous and correctly-pronounced words, when using acoustic scores of all phones*

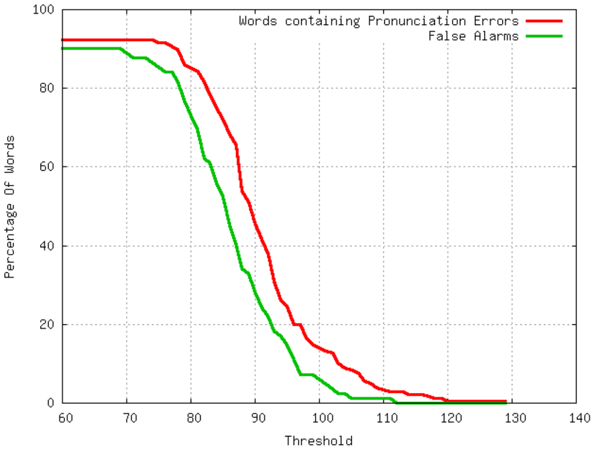

Figure 4: *Detection of erroneous and correctly-pronounced words, when using acoustic scores of vowels only*

of thresholds between 80 and 100.

With the introduction of an alternative dictionary containing the context dependent variations, a further improvement is observed. The variation selected from this extra dictionary directly influences the confidence level of a pronunciation decision made by the system. A simple implementation was used to generate the results shown in Figure 5 and Figure 6 with the introduction of a threshold penalty value. This value alters the average log likelihood of a phone with a fixed integer value. A phone is thus evaluated more leniently when the correct pronunciation of a word is selected form the alternative dictionary as well. The opposite is also true, resulting in the stringent evaluation of a phone when an alternative pronunciation of the word is selected from the alternative dictionary. Penalty values of 15 and larger cause saturation with our test data, but a penalty value of 10 produces a significant improvement in error detection.

To compare these three approaches, average values of the differences between the percentage of pronunciation errors detected and false alarms are calculated over a reasonable threshold interval. These values are shown in Table 2. An approvement of greater than 10% results as each technique is introduced.

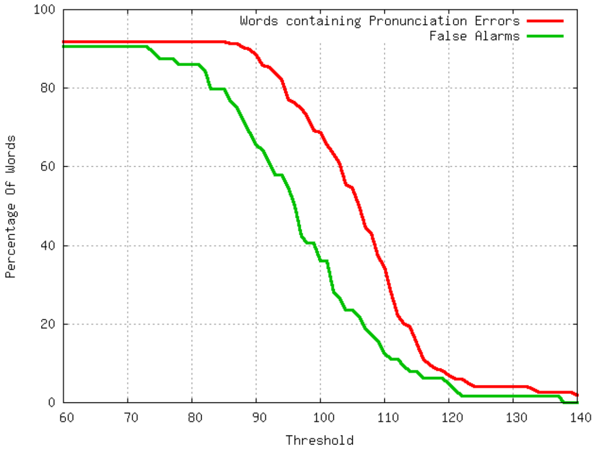

Figure 5: *Detection of erroneous and correctly-pronounced words, when using acoustic scores of vowels, and adding a threshold penalty of 5*

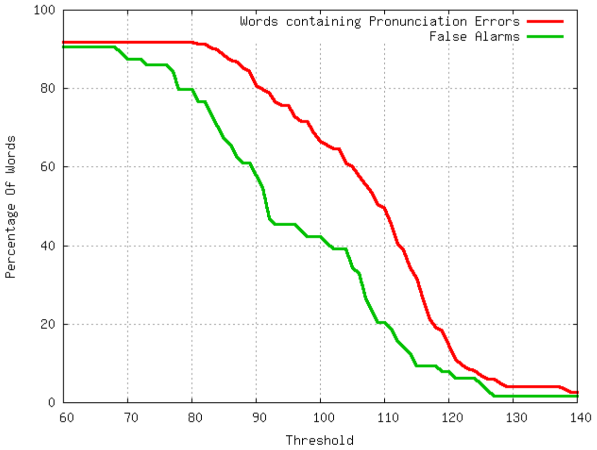

Figure 6: *Detection of erroneous and correctly-pronounced words, when using acoustic scores of vowels, and adding a threshold penalty of 5*

4.2. Tone classification

To obtain a measure of how well the system performs the automatic classification of tone levels on the syllable and word levels, a pre-labeled reference data set was used consisting of speech by a native isiZulu speaker. The reference set contained 23 sentence utterances comprising a total of 115 words and 435 syllables. It should be noted that the reference data was labeled subjectively by native isiZulu speakers independently from the acoustic data. Because the classification algorithm output depends on both the pitch and intensity of the speech, the system was tested with various pitch and intensity threshold parameters. Figures 7 and 8 show the results of the experiment on syllable and word levels. The different curves represent system evaluation at different intensity thresholds. Lower thresholds indicate that the classification is more readily affected by slight variations in the measured parameter. The points on the x-axis represent results at different *f0* thresholds. Another important

Technique	Threshold	Avg. Difference
NO	80 - 110	3.2%
VOWELS	80 - 110	12.1%
PENALTY 5	80 - 110	22.8%
PENALTY 10	80 - 110	24.8%

Table 2: *Pronunciation error detection performance*

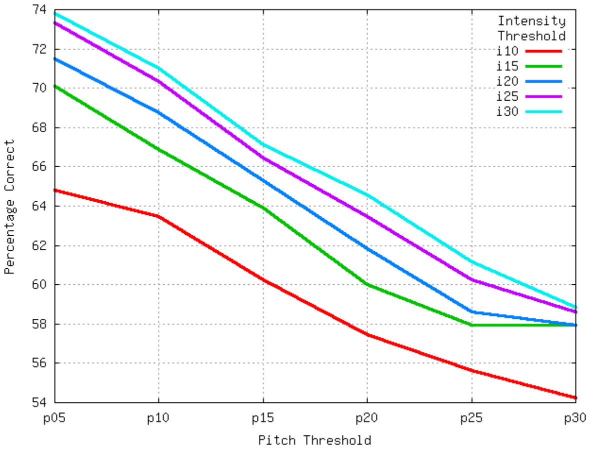

Figure 7: *Classification results (syllable level)*

point to note is that classification is done in two phases; firstly based on the *f0* characteristics and thereafter based on the signal intensity. Thus an initial classification is made according to pitch alone and states are subsequently re-evaluated according to the intensity criteria.

When interpreting the results presented, it is important to realize that the higher threshold values for a particular parameter effectively mean that the classification depends less on the specific parameter. We can thus infer from the relative positions of the different curves, that the classification is more precise if the intensity criterion is scaled down in significance. Similarly, it is evident from the slope of the curves that a system more sensitive in terms of variation in *f0* results in more accurate classification.

5. Conclusions

We have seen that the detection of phone pronunciation errors is practically possible by utilizing forced alignment output in conjunction with techniques exploiting specific language-dependent characteristics. From the results generated, it is evident that a substantial improvement in error classification is possible through these techniques. A viable system can be implemented by operating at the threshold that presents the best trade-off between pronunciation error recognition and false alarms. We believe that substantial additional improvements can be achieved by using acoustic models that more closely match the dialects that occur in the test environment, and are currently creating such models.

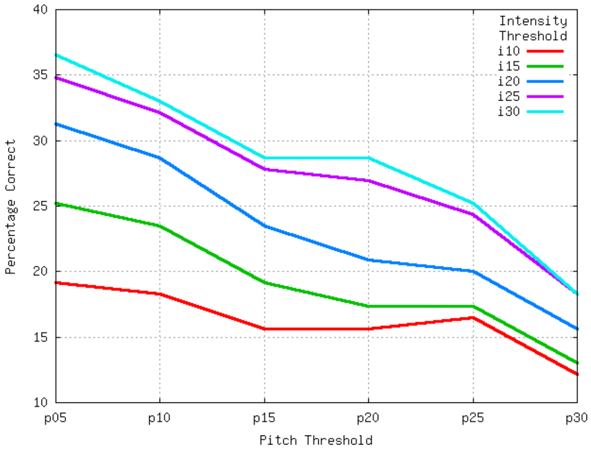

Figure 8: *Classification results (word level)*

We have also investigated the viability of an automatic tone classification system, and have obtained various measurements in order to assess implementation options and determine system relevance. The results highlight important classification parameters and their influence on system performance. Overall classification accuracy results support the viability of such a system. Further improvements will result from the use of additional training data, using phone boundary refinement techniques, possibly implementing pitch normalization before boundary determination and evaluating enhanced classification algorithms.

6. References

[1] A. Neri, C. Cucchiarini, and H. Strik, "ASR-based corrective feedback on pronunciation: does it really work?," in *Proceedings of Interspeech*, 2006, pp. 1982–1985.

[2] K. Probst, Y. Ke, and M. Eskenazi, "Enhancing foreign language tutors - In search of the golden speaker," *Speech Communication*, vol. 37, pp. 161–173, 2002.

[3] J.A. Louw, M. Davel, and E. Barnard, "A general-purpose IsiZulu speech synthesizer," *South African journal of African languages*, vol. 2, pp. 1–9, 2006.

[4] S. Paulo and L.C. Oliveira, *Advances in Natural Language Processing*, Springer Berlin / Heidelberg, http://www.l2f.inesc-id.pt, 2004.

[5] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Insitute of Phonetic Sciences*, 1993, vol. 17, pp. 97–110.

[6] K. Lee, "MLP-Based phone boundary refining for a TTS database," in *Proceedings of the IEEE*, May 2006, vol. 14, pp. 981–989.

[7] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Department, http://htk.eng.cam.ac.uk/, 2005.

[8] G. Kawai and K. Hirose, "Teaching the pronunciation of Japanese double-mora phonemes using speech recognition technology," *Speech Communication*, vol. 30, pp. 131–143, 2000.

[9] N. Govender, C. Kuun, V. Zimu, E. Barnard, and M. Davel, "Computational models of prosody in the Nguni languages," in *Proceedings of PRASA*, December 2005.

[10] N. Govender, *Intonation Modelling for the Nguni Languages*, M.Sc. Report, Dept. Computer Science, Pretoria Univ., 2006.

[11] P. Boersma, *Praat, a system for doing phonetics by computer*, Amsterdam: Glott International, 2001.