Acoustic analysis of diphthongs in Standard South African English

Olga Martirosian [1] and Marelie Davel [2]

[1] School of Electrical, Electronic and Computer Engineering,
North-West University, Potchefstroom, South Africa /
[2] Human Language Technologies Research Group, Meraka Institute, CSIR
omartirosian@csir.co.za, mdavel@csir.co.za

Abstract

Diphthongs typically form an integral part of the phone sets used in English ASR systems. Because diphthongs can be represented using smaller units (that are already part of the vowel system) this representation may be inefficient. We evaluate the need for diphthongs in a Standard South African English (SSAE) ASR system by replacing them with selected variants and analysing the system results. We define a systematic process to identify and evaluate replacement options for diphthongs and find that removing all diphthongs completely does not have a significant detrimental effect on the performance of the ASR system, even though the size of the phone set is reduced significantly. These results provide linguistic insights into the pronunciation of diphthongs in SSAE and simplifies further analysis of the acoustic properties of an SSAE ASR system.

1. Introduction

The pronunciation of a particular phoneme is influenced by various factors, including the anatomy of the speakers, whether they have speech impediments or disabilities, how they need to accommodate their listener, their accent, the dialect they are using, their mother tongue, the level of formality of their speech, the amount and importance of the information they are conveying, their environment (Lombard effect) and even their emotional state [1].

The nativity of a person's speech describes the combination of the effects of their mother tongue, the dialect that they are speaking, their accent and their proficiency in the language that they are speaking. If an automatic speech recognition (ASR) system uses speech and a lexicon associated with a certain nativity, non-native speech causes consistently poor system performance [2]. For every different dialect of a language, additional speech recordings are typically required, and lexicon adjustments may also be necessary.

Standard South African English (SSAE) is an English dialect which is influenced by three main South African English (SAE) variants: White SAE, Black SAE, Indian SAE and Cape Flats English. These names are ethnically motivated, but because each ethnicity is significantly related to a specific variant of SAE, they are seen as accurately descriptive [3]. Each variety will be made up of South African English as influenced specifically by the different languages and dialects thereof spoken in South Africa. It should be noted that these variants include extreme, strongly accented English variants that are not included in SSAE, and not referred to in this paper.

This analysis focuses on the use of diphthongs in SSAE. This is an interesting and challenging starting point to an acoustic analysis of SSAE. We are specifically interested in diphthongs since some of these sounds (such as /OY/ and /UA/, using ARPABET notation) are fairly rare and large corpora are required to include sufficient samples of these sounds.

A diphthong is a sound that begins with one vowel and ends with another. Because the transition between the vowels is smooth, it is modelled as a single phoneme. However, since it would also have been possible to construct a diphthong using smaller units that are already part of the vowel system, this may be an inefficient representation.

In this paper we evaluate the need for diphthongs in a lexicon by systematically replacing them with selected variants and analysing the system results. One way to analyse the phonemic variations in a speech corpus is to use an ASR system [4]. A detailed error analysis can be used to identify possible phonemic variations [1]. Once possible variations are identified, they can be filtered using forced alignment [4].

Some studies have found that using multiple pronunciations in a lexicon is better for system performance [5], while others have found that a single pronunciation lexicon outperforms a multiple pronunciation lexicon [6]. The argument can therefore be made for representing the frequent pronunciations in the data, but being careful not to over-customise the dictionary - if acoustic models are trained on transcriptions that are too accurate, they do not develop robustness to variation and therefore contribute to a decline in the recognition performance of the system [7].

In this paper we analyse diphthong necessity systematically in the context of an SSAE ASR system. The paper is structured as follows: In Section 2 we describe a general approach to identify possible replacement options for a specific diphthong, and to evaluate the effect of such replacement. In Section 3 we first perform a systematic analysis of four frequently occurring diphthongs individually, before replacing all diphthongs in a single experiment and reporting on results. Section 4 summarises our conclusions.

2. Approach

In this section we describe a general approach to first suggest alternatives for a specific diphthong and then to evaluate the effectiveness of these alternatives.

2.1. Automatic suggestion of variants

In order to identify possible alternatives (or variants) for a single diphthong, we propose the following process:

1. An ASR system is trained as described in more detail in Section 3.1.3. The system is trained using all the data available and a default dictionary containing the original diphthongs.

2. The default dictionary is expanded: variant pronunciations are added to words containing the diphthong in question by replacing the diphthong with all vowels and combinations of two vowels. Two glides (the sounds */W/* and */Y/*) are considered as part of the vowel set for the purpose of this experiment.

3. The original diphthong is removed completely, so that the dictionary only contains possible substitutions. The order of the substitutions is randomised in every word. This ensures that the speech that would represent the diphthong is not consistently labelled as one of the possible substitutions and the training process therefore biased in a certain direction.

4. The ASR system is used to force align the data using the options provided by the new dictionary. (Since the diphthong has been removed, the system now has to select the best of the alternatives that remain.)

5. The forced alignment using the expanded dictionary (alignment B) is compared to the forced alignment using the default dictionary (alignment A):

   - Each time the diphthong in question is found in alignment A, it and its surrounding phonemes are compared to the phonemes recognised at the same time interval in alignment B. The phonemes in alignment B that align with the diphthong in alignment A are noted as possible alternatives to the specific diphthong.
   - The alternatives are counted and sorted by order of frequency.

6. The frequency sorted list is perused and three to five possible replacements for the diphthong are selected by a human verifier from the top candidates. The human verifier is required to assist the system because they are equipped with SSAE and general linguistic knowledge, and are thus able to select replacement candidates that contain vowels or vowel combinations that are most likely to be replacements for the diphthong in question.

Once this process is completed, a list of possible replacements is produced. This list is based on a combination of system suggestion and human selection. For example, as a diphthong typically consists of two or more vowels linked together, it is quite likely that the best alternative to a diphthong is a combination of two vowels (diphone). Even though an ASR system may not initially lean towards such a double vowel replacement, including such an alternative may be forced by the human verifier. Also, knowledge-based linguistically motivated choices may be introduced at this stage. These choices are motivated by linguistic definitions of diphthongs as well as SAE variant definitions supplied in [3]. This process is described in more detail when discussing the process with regard to specific diphthongs below.

2.2. Evaluating replacement options

Once a list of three to five possible replacements has been selected for each diphthong, these replacements can be evaluated for their ability to replace the diphthong in question. Per diphthong, the following process is followed:

1. The default dictionary is expanded to include the selected alternatives as variants for the diphthong in question. The pronunciation with the diphthong is removed and the alternative pronunciations are randomised in order not to bias the system towards one pronunciation (as again, the system initially trains on the first occurring pronunciation of every word).

2. Each time the diphthong is replaced by an alternative, a list is kept of all words and pronunciations added.

3. An ASR system is trained on all the data using the expanded dictionary, and the alignments produced during training are analysed.

4. The pronunciations in the forced alignment are compared to each of the lists of added alternatives in turn, calculating the number of times the predicted pronunciation is used in the forced alignment, resulting in an occurrence percentage for each possible replacement.

5. Using these occurrence percentages, the top performing alternatives are selected. The number of selections is not specified, but rather, the ratio between the occurrence percentages of the alternatives is used to select the most appropriate candidates for the next round.

6. This process is repeated until only a single alternative remains, or no significant distinction can be made between two alternatives.

7. After each iteration of this process, the ASR phoneme and word accuracies are monitored.

3. Experimental Results

3.1. The baseline ASR system

In this section we define the baseline ASR system used in our experiments. We describe the dictionary used, the speech corpus and provide details with regard to system implementation.

3.1.1. Pronunciation Dictionary

The pronunciation dictionary consists of a combination of the British English Example Dictionary (BEEP) [8] and a supplementary pronunciation dictionary that has words contained in the speech corpus but not transcribed in BEEP. (This includes SAE specific words and names of places). The 44-phoneme BEEP ARPABET set is used. The dictionary was put through a verification process [9] but also manually verified to eliminate highly irregular pronunciations. The dictionary has 1 500 entries, 1 319 of which are unique words. The average number of pronunciations per word is 1.14 and the number of words with more than one pronunciation is 181. In further experimentation, this dictionary is referred to as the *default dictionary*.

3.1.2. Speech Corpus

The speech corpus consists of speech recorded using existing interactive voice response systems. The recordings consist of single words and short sentences. There are 19 259 recordings made from 7 329 telephone calls, each of which is expected to contain a different speaker. The sampling rate is 8 kHz and the total length of the calls is 9 hours and 2 minutes. It total, 1319 words are present in the corpus, but the corpus is rather specialised, with the top 20% of words making up over 90% of the corpus. For cross validation of the data, all the utterances of a single speaker were grouped in either the training or the test data, and not allowed to appear in both. The relevant phoneme counts are given in Table 1.

Table 1: *Selected phoneme counts for the speech corpus. Counts are calculated using forced alignment with the speech corpus and default dictionary. Diphthongs are shown in bold.*

Phoneme	Occurrences	Phoneme	Occurrences
/AX/	14 282	/UW/	3 151
/IY/	9 634	/AO/	3 106
/IH/	9 084	/Y/	2 743
/AY/	**6 561**	**/EA/**	**2 566**
/EH/	6 158	/ER/	2 499
/AE/	5 470	/AA/	2 097
/EY/	**4 509**	**/AW/**	**2 037**
/W/	4 293	/UH/	1 324
/AH/	3 883	**/IA/**	**1 014**
/OW/	**3 442**	**/UA/**	**455**
/OH/	3 232	**/OY/**	**39**

3.1.3. System Particulars

A fairly standard ASR implementation is used: context dependent triphone acoustic models, trained using Cepstral Mean Normalised 39-dimensional MFCCs. The optimal number of Gaussian Mixtures per state in the acoustic models was experimentally determined to be 8. The system makes use of a flat word based language model and was optimised to achieve a baseline phoneme accuracy of 79.57% and a corresponding word accuracy of 64.50%. As a measure of statistical significance, the standard deviation of the mean is calculated across the 10 cross-validations, resulting in 0.07% and 0.13% for phoneme and word accuracy respectively. The system was implemented using the ASR-Builder software [10].

3.2. Systematic replacement of individual diphthongs

In this section we provide results when analysing a number of diphthongs individually according to the process described in the previous section (Section 2).

Since training the full system outlined in Section 3.1.3 is highly time consuming, a first experiment was performed to determine whether a monophone-based system is sufficient to use during the process to identify and evaluate replacement options. For each diphthong investigated, a dictionary was compiled as described in Section 2.1, a full system was trained using this dictionary, and its forced alignment output when using monophone models was compared with its forced alignment output when using triphone models with 8 mixtures. This comparison always resulted in an equivalence of more than 95%. Therefore, from here onwards, only monophone alignment is used for decision making, while final accuracies, or selection rates, are reported on using the full triphone system.

3.2.1. Diphthong Analysis: /AY/

The AY diphthong was first to be analysed. The results of the analysis are summarised in Table 2. Each line represents one experiment. For each experiment, the accuracies of each of the included alternatives are noted, as well as the cross validated phoneme and word accuracies of the full ASR system.

The progression of this experiment is outlined below:

- In the first iteration, the alternatives /AH/, /AH IH/ and /AA/ achieve the highest accuracies and are selected for the next round. /AH/ achieves the highest selection rate overall.
- In the second iteration, the alternatives /AH/ and /AA/ achieve the highest accuracies and are selected for the next round. Again, /AH/ has the highest selection rate. All diphones have now been eliminated.
- In the third iteration, /AH/ has the highest selection rate and is therefore selected as the final and best alternative for /AY/.
- In the fourth iteration, /AH/ is tested as a replacement of /AY/. Phoneme accuracy rises to its highest, however, word accuracy suffers. As phoneme accuracy in influenced by the change in number of phonemes (from one experiment to another), word accuracy is the more reliable measure for this experiment.
- The diphone theory, detailed in Section 2.1, suggests that, because diphthongs are made up of two sounds, their replacement must also consist of two sounds in order to have the capacity to model them accurately. In order to test this theory, an iteration is run with /AH/ and /AH IH/ as the alternatives for /AY/. The ASR system still selects the /AH/ alternative over the /AH IH/ alternative. However, the word accuracy increases at this iteration, implying that perhaps having /AH IH/ as an alternative pronunciation for /AY/ fits the acoustic data better than only having /AH/.
- A final iteration is run with the knowledge-based linguistically motivated choice "/AH IH/" as the replacement of /AY/. Both the phoneme and word accuracy rise to their highest values with this replacement. This shows that the linguistically predicted /AH IH/ is indeed the best replacement for /AY/.

3.2.2. Diphthong Analysis: /EY/

The /EY/ diphthong is analysed using the technique outlined in Section 2. The results are summarised in Table 3. In the first iteration, /AE/ and /EH/ are clearly the better candidates, but the diphone (double vowel) scores were lower and very similar. Thus, for the second iteration, all diphones are cut and only /AE/ and /EH/ are tested. But for the third iteration, testing the necessity of including a diphone, two of the diphones were brought back to be tested again. It should be noted that the highest word accuracy achieved for the suggested variants was achieved in the third iteration, suggesting that diphones are indeed necessary when attempting to replace a diphthong. Again, the highest accuracy achieved overall is for the knowledge-based linguistically suggested alternative /EH IH/.

3.2.3. Diphthong Analysis: /EA/

The /EA/ diphthong is now analysed. The results of the experiment are summarised in Table 5. These results behave quite differently compared to the other diphthong experiments. The first iteration, where all 3 of the variant options are included, achieves the highest word accuracy, even higher than the iteration which makes use of linguistic knowledge. The phoneme accuracy however, increases with every iteration, reaching its peak with the use of the linguistic replacement. Again, this may be related to the change in number of phones (in words causing errors) which makes word accuracy a more reliable measure. The knowledge-based linguistic replacement performs very well, achieving the second highest word accuracy overall.

Table 2: *Results of the experiments for the diphthong /AY/*

	/AH/	/AA/	/AH IH/	/AE IY/	/AH IY/	P Acc	W Acc
1	0.46	0.20	0.18	0.08	0.07	78.51%	63.88%
2	0.46	0.36	0.17	N/A	N/A	78.75%	64.06%
3	0.56	0.43	N/A	N/A	N/A	79.14%	64.17%
4	1	N/A	N/A	N/A	N/A	79.56%	64.03%
5	0.62	N/A	0.38	N/A	N/A	79.19%	64.13%
6	N/A	N/A	1	N/A	N/A	79.77%	64.30%

Table 3: *Results of the experiments for the diphthong /EY/*

	/AE/	/EH/	/AE IY/	/AE IH/	/EH IY/	/EH IH/	P Acc	W Acc
1	0.24	0.25	0.17	0.17	0.16	N/A	78.97%	64.27%
2	0.59	0.41	N/A	N/A	N/A	N/A	79.30%	64.03%
3	0.48	N/A	0.26	0.27	N/A	N/A	79.36%	64.41%
4	1	N/A	N/A	N/A	N/A	N/A	79.64%	64.04%
5	N/A	N/A	N/A	N/A	N/A	1	79.78%	64.43%

Table 4: *Results of the experiments for the diphthong /OW/*

	/OH/	/ER/	/ER UW/	/AE/	/AE UW/	/AX UH/	P Acc	W Acc
1	0.29	0.36	0.14	0.13	0.08	N/A	79.53%	64.33%
2	0.52	0.48	N/A	N/A	N/A	N/A	79.57%	64.41%
3	0.59	N/A	0.41	N/A	N/A	N/A	79.53%	64.48%
4	1	N/A	N/A	N/A	N/A	N/A	79.60%	64.45%
5	N/A	N/A	N/A	N/A	N/A	1	79.63%	64.48%

Table 5: *Results of the experiments for the diphthong /EA/*

	/EH/	/IH EH/	/AE/	/EH AX/	P Acc	W Acc
1	0.51	0.34	0.15	N/A	79.22%	64.49%
2	0.72	0.28	N/A	N/A	79.51%	64.43%
3	1	N/A	N/A	N/A	79.65%	64.21%
4	N/A	N/A	N/A	1	79.73%	64.30%

Table 6: *IPA based diphthong replacements*

Diphthong	Diphone	Diphthong	Diphone
/AY/	/AH IH/	**/OY/**	/OH IH/
/EY/	/EH IH/	**/AW/**	/AH UH/
/EA/	/EH AX/	**/IA/**	/IH AX/
/OW/	/AX UH/	**/UA/**	/UH AX/

3.2.4. Diphthong Analysis: /OW/

The experiment is repeated for the diphthong /OW/. The results for the experiment are outlined in Table 4. The phoneme accuracy follows a similar pattern to the earlier experiments. The word accuracy is highest at both iteration 3, where a diphone is included and iteration 5, where the linguistic knowledge-based replacement is implemented. The knowledge-based linguistic replacement once again achieves the highest phoneme and word accuracies.

3.3. Systematic replacement of all diphthongs

Given the results achieved in the earlier experiments, a final experiment is run where all the diphthongs are replaced using a systematic system based on the linguistic definitions of the individual diphthongs.

Two ASR systems are used, designed as described in Section 3.1.3. These two systems differ only with regard to their dictionary. One system (system A) uses the baseline dictionary, in the other (system B), the diphthongs in the baseline dictionary are all replaced with their diphone definitions, using British English definitions defined in Table 6.

All results are cross-validated and the two systems are compared using their word accuracies. Interestingly word accuracy decreases only very slightly: from 64.53% for system A to 64.35% for system B. The removal of 8 diphthongs is therefore not harmful to the accuracy of the system. This is an interesting result, especially as the detailed analysis was only performed for 4 of the diphthongs and further optimisation may be possible.

4. Discussion

The aim of this study was to gain insight into the use of diphthongs in SSAE. We defined a data-driven process through which diphthongs could automatically be replaced with optimal phonemes or phoneme combinations. To complement this process, a knowledge-based experiment was set up using linguistic data for British English. Although the data-driven method was partially successful in finding the best replacement for diphthongs, the knowledge-based method was superior. However, the increase in accuracy from the knowledge-based method is small enough that if knowledge is not available, the data-driven technique can be used quite effectively.

It is interesting to consider the South African English variants that are described in [3]. The variants described here or ones close to them always appear on the list of the top candidates of the data-driven selection. This in itself is an interesting observation from a linguistic perspective.

From a linguistic perspective, the fact that a diphthong can successfully be modelled as separate phonemes provides an insight into SSAE pronunciation.

From a technical perspective, the removal of diphthongs simplifies further analysis of SSAE vowels. Our initial investigations were complicated by the confusability between diphthongs and vowel pairs, and this effect can now be circumvented without compromising the precision of the results.

Ongoing research includes further analysis of SSAE phonemes with the aim to craft a pronunciation lexicon better suited to South African English (in comparison with the British or American versions commonly available). In addition, similar techniques will be used to evaluate the importance of other types of phonemes, for example the large number of affricates in Bantu language.

5. References

[1] Strik H. and Cucchiarini C., "Modeling pronunciation variation in ASR: A survey of the literature," *Speech Communication*, vol. 29, pp. 225–246, 1999.

[2] Wang Z., Schultz T., and Waibel A., "Comparison of acoustic model adaptation techniques of non-native speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Hong Kong, April 2003, vol. 1, pp. 540 – 543.

[3] Kortmann B. and Schneider E.W., *A Handbook of Varieties of English*, vol. 1, Mouton de Gruyter New York, 2004.

[4] Adda-Decker M. and Lamel L., "Pronunciation variants across system configuration, language and speaking style," *Speech Communication*, vol. 29, pp. 83–98, 1999.

[5] Wester M., Kessens J.M., and Strik H., "Improving the performance of a dutch csr by modelling pronunciation variation," in *Proceedings of the Workshop Modeling Pronunciation Variation for Automatic Speech Recognition*, Rolduc, The Netherlands, May 1998, pp. 145–150.

[6] Hain T., "Implicit modelling of pronunciation variation in automatic speech recognition," *Speech communication*, vol. 46, no. 2, pp. 171–188, 2005.

[7] Saraclar M., Nock H., and Khudanpur S., "Pronunciation modeling by sharing gaussian densities across phonetic models," in *Sixth European Conference on Speech Communication and Technology*, Budapest, Hungary, September 1999, ISCA.

[8] BEEP, "The british english example pronunciation (beep) dictionary," `ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries`.

[9] Martirosian O.M. and Davel M., "Error analysis of a public domain pronunciation dictionary," in *PRASA 2007: Eighteenth Annual Symposium of the Pattern Recognition Association of South Africa*, Pietermaritzburg, South Africa, November 2007, pp. 13–18.

[10] Mark Zsilavecz, "ASR-Builder," January 2008, `http://sourceforge.net/projects/asr-builder`.