Vowel Quality in Accent Modification

Jonathan Teutenberg, Catherine Watson

Department of Computer Science; Department of Electrical and Computer Engineering
University of Auckland, New Zealand
jono@cs.auckland.ac.nz; c.watson@auckland.ac.nz

Abstract

The contribution of a change in vowel quality to the perception of accent in synthesised speech is assessed. An NZE voice is created against which a baseline transformation is compared. Listening tests are performed to assess the success of the baseline transformation at modifying the perceived accent of the transformed speech. It is shown that even a simple transformation can yield a significant shift in perceived accent. However, it also appears that vowel quality alone is insufficient for the modification of accent, and that a successful vowel quality component of an accent modification system can reach no more than 3.8 on an MOS-like scale.

1. Introduction

The localisation of synthesised speech systems is usually performed through the construction of a wholly new synthetic voice using the target accent. An alternative to this often time consuming process could be to apply an accent transformation to speech produced by an existing voice. The output of such a transformation should be speech that is as intelligible as the original, but is perceived as being spoken in the target (local) accent.

Accent is taken to be comprised of the sub-lexical features of speakers that contribute to the distinction between groups of speakers by geographic, cultural or socio-economic factors. Such features include pronunciation and prosody, but exclude word choice and other high-level features.

In this study we focus on a very simple accent transformation, and only consider those features relating to vowel quality. The term vowel quality refers to the positions of articulators such as lips, tongue and mouth during the production of the sound, which affect frequencies of the formants that in turn affect the classification of the phone by a listener. The first and second formants have the most significant contribution to the quality of vowels, and points on an F1/F2 plane are often used to represent a speaker's pronunciation. The collection of points in the F1/F2 plane given by a speaker's cardinal vowels is referred to as their vowel space. In this paper we shall refer to these vowels using Wells' keyvowel system, designed to give full coverage of the vowel spaces of the accents of English (Wells 1982).

Past work has considered a more complete transformation of the formants of natural Australian and British speech (Yan & Vaseghi 2003). Perceptual tests indicated that modification of the first five formants has some impact on the perceived accent (Yan, Vaseghi, Rentzos, & Ho 2004). Four important differences between previous work and ours are: one, that we are assessing synthetic rather than natural speech which introduces significant noise; second, the focus on only the first two formants of speech; thirdly, that we avoid introducing artifacts from signal processing and finally that we obtain statistically significant results.

We discuss the new New Zealand English (NZE) voice we have developed and present some preliminary results towards the creation of an accent transformation system. We compare a voice using phones recorded in our target accent (NZE) with one obtained by re-mapping the vowel space of a British Received Pronunciation (RP) speaker.

2. NZE voice

We have created a NZE voice for the festival speech synthesis system, based on diphones. It consists of a full range of around 2000 diphones of an adult male speaker recorded in a quiet, dampened room over 8 hours split into four sessions held at the same time each day. One example of each diphone was recorded placed within a nonsense word as per the standard methods specified in the festvox documentation (Black & Lenzo 2000). After recording, the nonsense words were automatically labelled using the Edinburgh Speech Tools (Taylor, Caley, & Black 1998) followed by some hand correction of diphone boundaries. A diphone voice was chosen over one based on unit selection as we intend to include prosody modification in later experiments. At present unit selection voices, despite giving significantly higher quality speech, allow limited control over the prosodic features of the synthesised speech (Clark & King 2006).

The phones used are those that appear in the NZE accented UNISYN lexicon (Fitt & Isard 1999). These include phones for the tapped 't', merged 'ear' and 'air' diphthongs, and the dark 'l'. The potential inclusion of Maori names and phrases is considered, so utterance-final short vowels are also included in the database even though they do not occur in English.

For this study the NZE voice is based on entries in the same pronunciation dictionary (OALD) and uses the same prosody as the RP voice in Festival, the rab diphone voice. Thus the voices differ only in their pronunciation of the same phones. The intention here is to isolate the contribution of vowel quality to perceived accent from that of lexical differences.

3. Mapping vowel spaces

A two dimensional space spanned by the first and second formants is often used as a simplification of the articu-

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

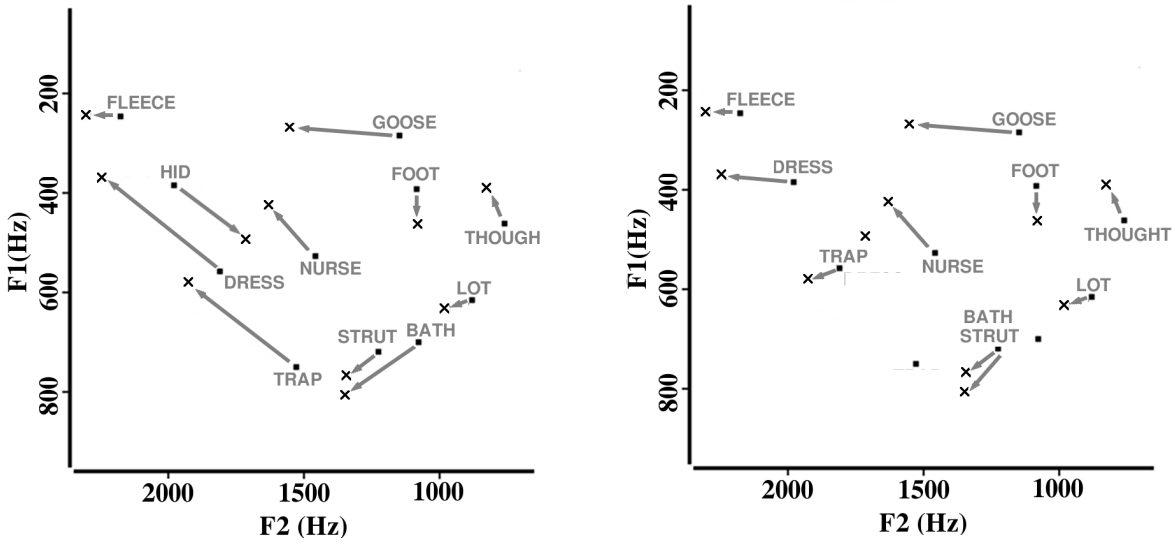

Figure 1: Difference between RP and NZE vowel spaces (left) and the difference between remapped-RP and NZE (right).

latory features that make up vowel quality. The F1 dimension being correlated to the openness of the jaw and the F2 dimension to the frontedness of the tongue during production.

Since vowels that are close in F1/F2 plane are also perceptually close, the overall distance between a phonemically identical pair of vowels of two speakers can be used as a measure of the difference in their pronunciations. In terms of accent transformation, we would like to minimise the distances between the vowel spaces of the source speaker and a typical (or mean) speaker of the target accent.

While NZE and RP are easily discernable to local speakers, they contain many similarities. For example they have the same number of monophthongs, and both are non-rhotic. It is therefore possible that by substituting the RP vowel closest to the the various NZE phones in the NZE vowel space within the F1/F2 plane it will sufficient to allow a 'remapped' RP voice to be perceived as NZE. For instance we represent the NZE TRAP vowel with the RP DRESS vowel. Figure 3. shows the distances between pairs of vowel spaces of NZE (Watson, Harrington, & Evans 1998) and RP (Deterding 1990). The left shows the difference between a typical NZE speaker and an RP speaker, and the right shows NZE and remapped RP vowels. The length of each arrow gives the distance between the vowel in each accent, with the direction toward the NZE pronunciation in both diagrams. The labels show which vowel the arrow refers to. Thus it can be seen that after remapping the RP vowel space toward NZE, the BATH and STRUT vowels have now merged and both will use the RP BATH whenever they occur in an utterance. It is clear that, at least visually, the remapping results in a significant reduction in the overall difference between the two vowel spaces. A reduced distance between the vowel spaces is expected to result in a perceived accent closer to the target.

We acknowledge that this will never be as accurate a transformation as can be performed with signal processing, however determining how effective this remapping is in a perceptual sense is of interest. It is also plausible that the highly distinctive vowels act as 'flags' for classification of accent, in which case this method may be quite effective as most of the vowels with large distance between accents are to be altered.

Text (A)	Russia s**a**ys **i**t w**i**ll restore g**a**s to Europe
IPA	rʌʃə sɛz ɪt wɪl rɪstɔː gas tuː jʊərəp
Remapped	rʌʃə sɪz ət wəl rɪstɔː gɛs tuː jʊərəp
Text (B)	**Se**veral c**a**rs h**a**ve cr**a**shed h**ere** **i**n the p**a**st
IPA	sɛvrəl kɑːz hav kraʃt hɪə ɪn ðə pɑːst
Remapped	sɪvrəl kʌz hɛv krɛʃt hɪə ən ðə pʌst
Text (C)	The sh**a**rks don't really l**i**ke c**a**tfish
IPA	ðə ʃɑːks dəʊnt rɪəli laɪk katfəʃ
Remapped	ðə ʃʌks dəʊnt rɪəli leɪk kɛtfəʃ
Text (D)	W**i**tnesses s**a**y sn**a**ke **a**te bear
IPA	wɪtnəsəz seɪ sneɪk ɛt beə
Remapped	wətnəsəz saɪ snaɪk ɪt beə

Table 1: The four sentences used for perceptual tests, with IPA transcriptions before and after remapping.

4. Method

4.1. Sentences

Four sentences were synthesised for the perceptual tests. The sentences were chosen so as to cover a wide range of vowels eligible for remapping, as well as a mixture of vowels that remained unchanged. Table 4. gives a full listing of the sentences and their mapped and unmapped IPA transcription. Between 33% and 78% of vowels are eligible for remapping.

Sentences B and C have utterance final words containing remapped vowels, whereas A and D do not. Should the perception of accent be dependent on the presence of a

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

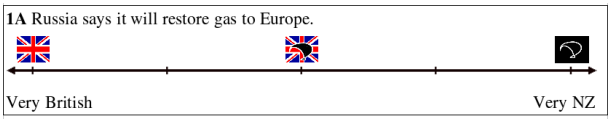

Figure 2: A typical panel on which listeners place the accent of a voice for one sentence.

single distinctive phone we can expect this to show up in a comparison between these sentences. Sentence C is of particular interest as it contains the TRAP vowel which undergoes a significant change in position under remapping and contains few other altered vowels.

4.2. Voices

Each sentence was synthesised by the festival framework using four voices: the original RP voice, the remapped RP voice, the NZE voice and an NZE voice with pronunciation shifted toward RP via spectral warping. The fourth voice was included in response to problems raised in preliminary surveys as described in Section 4.3..

4.3. Perceptual tests

8 adult native New Zealand speakers unfamiliar with synthesised speech were asked to listen to the four versions of these sentences and to place the accent on a scale from "Very British" to "Very NZ". In addition to text labels, three easily recognisable icons were placed on the scale as shown in figure 4.3.. It was felt that these icons gave a better impression of a continuum on which to place the accents rather than forcing participants to select from a given set of categories. The participants were also provided with the text of the sentence prior to hearing.

Preliminary surveys showed that participants were adverse to placing different voices in the same category, even if the perceived accent was very similar, hence the use of a continuous scale. Other problems include listeners recognising the 'correct' NZE voice from differences in voice quality rather than pronunciation and associating each of three voices with one of the three icons on the continuum. While analytically features of accents tend not to lie on a continuum, perceptually they do. For example, while Southern and Northern American English have two distinct treatments of the diphthong /aɪ/, artificially constructed versions of this diphthong were perceived to belong to accents of states on a continuum from north to south, even though no such continuum exists (Clopper & Pisoni 2004). We feel this is justification for not simply asking listeners to label a voice as either British or New Zealand.

The sample size was determined by simply increasing the number of listeners polled until the results reached a high level of confidence.

The sentences were presented in order from A to D, with the order of the four voices randomised for each sentence. One exception was the very first voice heard, which was set as the additional frequency warped voice. This allowed the listeners to become accustomed to the sound of the synthetic voices without affecting our results.

	Sentence			
Speaker	A	B	C	D
RP	1.8(0.5)	2.0(0.2)	1.7(0.5)	2.1(0.7)
Remapped	2.9(0.6)	3.2(0.9)	2.4(0.4)	3.1(1.0)
NZE	4.1(0.5)	3.7(0.7)	3.7(0.7)	3.9(0.8)

Table 2: Mean scores and standard deviations for 'NZEness' of RP, remapped RP and NZE voices for each sentence

Each participant was provided the survey in a reasonably quiet office environment on the same equipment. Each voice on each sentence was played once, with the exception of the very first which was played around three times so that the participant felt comfortable with the synthetic speech. The total time taken per participant was less then 10 minutes in all cases.

5. Results and discussion

The perceived accents were given a MOS-like value between 1 (Very British) and 5 (Very NZ), but were not quantised. Table 5. shows the mean scores over all listeners for each sentence. The remapped RP voice is significantly more 'NZ' than the original RP voice with 99% confidence for the first three sentences by a Student's t-test. On sentence D the remapped voice was more NZ than the original with only 96% confidence due to a higher variance in listener responses. This variance is most likely due to a slightly inaccurate remapping of the FACE diphthong to PRICE which caused some participants to add comments such as "Sounds Australian!" (mentioned by two of the participants).

The trend of the perception of the remapped voice as NZ over the first three sentences follows the percentage of remapped vowels. Thus sentence B with 78% of its vowels remapped was perceived most strongly as NZ, and the remapping of TRAP in a sentence-final position in sentence C had no apparent effect beyond that of other remapped vowels.

In most cases the NZE voice was perceived as significantly more NZ than the remapped RP voice with high confidence. However on sentence B, where the remapping was most effective, a t-test can only claim 77% confidence. While taking a larger sample would most likely show a more definite difference between the two voices this result shows the effectiveness of this very simple transformation.

Speaker	Score (NZEness)
RP	1.9(0.3)
Remapped	2.7(0.5)
NZE	3.8(0.5)

Table 3: Mean scores and standard deviations for 'NZEness' of RP, remapped RP and NZE voices over all sentences

Table 5. shows the results for each voice aggregated over all sentences. Despite the small sample size, the results show significant differences between RP and remapped RP,

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review

and between remapped RP and NZE. Clearly the remapping is a useful technique but is by no means a solution to the problem of vowel quality modification.

A recording of a native NZ speaker over the diphones selected by a RP voice should be at least as 'NZ' as any method of transforming pronunciation. Thus on a scale from 1 to 5 on RP to NZE conversion any transformation method scoring 3.8 in these circumstances should be considered successful. Ideally the perceptual tests would include a second example voice for both RP and NZE against which participants would rate the various transformed versions. Unfortunately synthetic voices are few, and to our knowledge no other NZE voice yet exists that could be used as a comparison.

An expected application of the remapping of a vowel space was to reduce the distance to a target vowel space prior to the application of signal processing based transformations. The assumption here is that adjusting formants over a shorter distance produces fewer artifacts, and that the remapping itself effectively produces no artifacts thus resulting in higher quality transformed speech. In practice, however, we have found that some artifacts are produced during remapping. For example, when the long BATH vowel is replaced by the short vowel STRUT. Here a concatenative synthesiser must stretch the short STRUT from its database to fit the longer duration meant for BATH, producing perceptible artifacts. Restricting the remapping to long-to-long and short-to-short vowels would remove this problem, but would also cripple the potential reduction in distance to the target vowel space. The remapping is probably best left as a useful baseline comparison rather than a part of future accent modification systems.

References

Black, A. & K. Lenzo (2000). Building Voices in the Festival Speech Synthesis System. *Retrieved on 11/05 from http://festvox.org*.

Clark, R. & S. King (2006). Joint Prosodic and Segmental Unit Selection Speech Synthesis. In *Proceedings of INTERSPEECH*.

Clopper, C. & D. Pisoni (2004). Some Acoustic Cues for the Perceptual Categorization of American English Regional Dialects. *Journal of Phonetics 32*, 111–140.

Deterding, D. (1990). Speaker normalisation for automatic speech recognition, phd thesis. In *Gimson's Pronunciation of English, 5th Edition, Revised by Alan Cruttenden.*

Fitt, S. & S. Isard (1999). Synthesis of Regional English Using a Keyword Lexicon. In *Proceedings of Eurospeech 99*, Volume 2, pp. 823–826.

Taylor, P., R. Caley, & A. Black (1998). The Edinburgh Speech Tools Library. *Retrieved from 11/05 from http://www.cstr.ed.ac.uk/projects/speech_tools/*.

Watson, C., J. Harrington, & Z. Evans (1998). An acoustic comparison between New Zealand, and Australian English vowels. *Australian Journal of Linguistics 18*(2), 185–207.

Wells, J. (1982). *Accents of English*. Cambridge University Press.

Yan, Q. & S. Vaseghi (2003). Analysis, modelling and synthesis of formants of British, American and Australian accents. In *International Conference on Acoustics, Speech and Signal Processing Proceedings*, Volume 1, pp. 712–715.

Yan, Q., S. Vaseghi, D. Rentzos, & C. Ho (2004). Analysis by Synthesis of Acoustic Correlates of British, Australian and American Accents. In *International Conference on Acoustics, Speech and Signal Processing Proceedings*, Volume 1, pp. 637–640.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after abstract only review