The Effect of Audience Familiarity on the Perception of Modified Accent

Jonathan Teutenberg[1], *Catherine I. Watson*[2]

[1]School of Computing, Teesside University, United Kingdom
[2]Department of Electrical and Computer Engineering, University of Auckland, New Zealand
j.teutenberg@tees.ac.uk, c.watson@auckland.ac.nz

Abstract

Evaluating the efficacy of accent transformation is important when localising speech-enabled software. However, perceived accent is an attribute assigned by a listener, and the apparent success of accent transformation will vary with the audience. Here we show the extent to which evaluations can be affected by audience familiarity with an accent. A perceptual study comparing two approaches to accent transformation is presented to two audiences with differing familiarity with the target accents. For mean opinion score style evaluations, we quantify the approximate change in perception, and show that this can be sufficient to alter relative successfulness of such systems.
Index Terms: accent perception, voice conversion, localisation, evaluation

10.21437/Interspeech.2010-559

1. Introduction

Producing new synthetic voices for speech synthesis systems is expensive in time, and requires expert knowledge. This has motivated research into voice conversion–automatically adapting a source speaker into a target speaker based on a small corpus of examples. An accent transformation system describes a single mapping that can be applied to speech from any voice in a source accent to produce similar speech in a target accent, and can be seen as a variant of voice conversion. This approach to adapting speech is particularly appropriate for localisation tasks, and unlike generic voice conversion, it does not suffer from the requirement of training corpora for every source voice used. However, the criteria for the evaluation of accent transformation systems requires unique attention.

The most obvious application for the transformation of accent is to adopt an audience's accent so as to build rapport [1]–an application that is garnering increased interest in synthetic speech production [2, 3]. However, other equally useful applications exist in which the target audiences may not always be overly familiar the desired target accent [4]. Listeners identify a population of individuals with an accent of speech, and by association the individual by whom that speech was produced. Thus by accent alone, an individual can assume properties of their audience's conceptual view of a population such as social status, level of education, or affluence [5, 6].

When an audience is less familiar with a target accent, their perception of it relies on only a subset of the accent's features. Evidence for this has been shown in a number of studies [7, 8, 9] in which listeners with knowledge of the accent presented (in this case northern US residents) were able to accurately associate a speaker with their geographic location (specific northern state). However nãive listeners (from southern states) were only able to distinguish north from south with any accuracy. Thus finer details of accent differences had no perceptual effect for the unfamiliar audience. Further investigations [10] have found that the number of acoustic features that are used by nãive listeners when classifying accents is, in general, around two. The features assessed were vowels' location in the vowel space, vowel 'brightness' and 'r-fulness', fricative voicing and monophthongisation. So, for example, an occurance of a backed realisation of a /u:/ phoneme alone was seen to induce a not-southern classification (the region where /u:/ is usually fronted). It should be noted that these salient features of accent are not restricted to the phonetic level of vowel pronunciation, but can also include intonation and prosodic features [11].

Therefore the apparent efficacy of an accent transformation system can be expected to vary significantly depending on the audience used for evaluation. In this paper we attempt to determine the extent to which this audience familiarity can be expected to impact on the results of evaluations. We wish to determine whether, in the comparison of two or more accent transformation systems, an apparently more successful transformation as perceived by one audience, can be perceived as the less successful approach by another audience. We also wish to quantify the expected differences in perceived accent when using numeric ratings such as mean opinion scores, between audiences that are familiar and unfamiliar with the target accent.

To answer these questions, we run two perceptual studies assessing the efficacy of a pair of accent transformation approaches. Speech samples include transformations to UK accents and antipodean accents from British received pronunciation (RP). The two studies present the same speech samples to two different audiences: firstly listeners from north-east England, and secondly of listeners from New Zealand.

2. Accent Transformation Approach

In this section we give an overview of the approach taken to transform the accent of natural utterances. As this paper focuses on the relative differences in audience perception of the resulting speech rather than the approach itself, in the interest of brevity the details of the approach have been abridged. Examples of the transformed speech samples used in the evaluations discussed in this paper can be found at *http://www.cs.auckland.ac.nz/~jteu004/Interspeech2010/index.html* which may more readily provide an understanding of the transformations being applied.

The modifications used to transform the accent were made primarily to pronunciation and fundamental frequency. Natural utterances of speech were analysed with a harmonic/stochastic-based approach, modified, and resynthesised using modulated sinusoids. This approach is a hybrid of HNM_1 [12] and the spectral representation used by STRAIGHT [13]. A full description of the resynthess approach is given in [14]. While the annotation of input data (such as phone boundaries and pitch marks) was checked by hand, the remainder of the accent transforma-

tion process is an automated procedure.

In prior perceptual studies we have shown this approach to be effective at transforming perceived accent of natural speech [15].

2.1. Pronunciation Transformation

Modifications to pronunciation were realised by shifting formants. Most of these changes were made on the first two formants, positioning them within an F_1/F_2 space, such that minimum and maximum formant frequencies for each speaker were preserved (accounting somewhat for vocal tract size). The frequencies of higher formants were also adjusted, particularly when producing rhotic phones from non-rhotic vowel phonemes. For analysis frame, the formants of the spectral envelope were modelled as a sum of independent exponential functions. One exponential function was used per formant, plus one to model the spike at the fundamental frequency that occurs when using harmonic/stochastic analysis. A piece-wise linear function was used to represent spectral slope. Formants were shifted by adjusting the centre of each exponential function, and modifying their magnitude according to the value of the spectral slope at the target frequency. A target frequency was determined for each formant on each phone, based on Wells' collection of phonetic descriptions of English accents [16], and a mapping from the set of International Phonetic Alphabet phones to a normalised F_1/F_2 space. With formant targets specified, temporal changes in the formant frequencies–the phone transitions–were modelled with half-cosines.

2.2. Prosodic Transformation

Fundamental frequency contours were modified by imposing natural contours taken from similar phrases in the target accent. F_0 contours were represented using a two-tiered Discrete Cosine Transform (DCT) approach, based on [17]. The two tiers of a contour were represented by a set of DCT coefficients at the phrase level, and a set of coefficients for each syllable. Dynamic time warping was applied at analysis time, so that the contours' syllable durations were normalised. This normalisation was performed to reduce the dependencies between the fundamental frequency representation and phone durations or local speech rate.

Limited duration modifications were made. This was restricted to those phonemes that are realised by a long phone in the source accent and a short phone in the target accent (or vice-versa) according to Wells. An example of this is when transforming the /a:/ phoneme from the long British RP [ɑ̟ː] phone to the short Irish [a] phone.

3. Experimental Setup

3.1. Speech Data

Utterances of the North Wind story read by male and female RP speakers were taken from the Keele speech corpus [18]. The story was split into nine utterances, each of which was associated with a randomly selected speaker (five male and four female) as shown in Table 1. Each of these nine source utterances were then associated with one or more target accents from: Australian, New Zealand, Irish and Welsh. A northern Irish target pronunciation was used for the "Irish" accent. So as to limit the duration of surveys, source-target pairs were chosen so that each target accent appeared three times, and it was ensured that each source utterance appeared at least once.

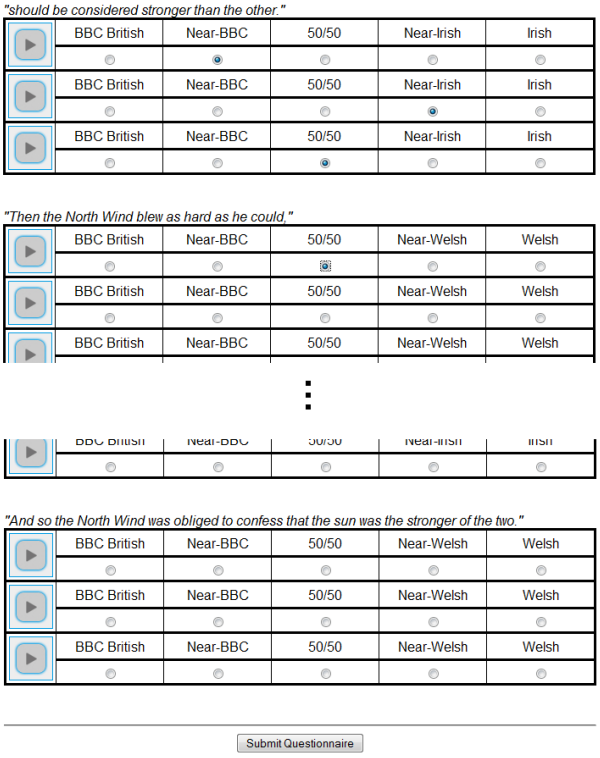

Figure 1: *Survey format used in the evaluations.*

For each source-target pair, two transformed utterances were generated. The first performed only pronunciation modification, and the second performed both pronunciation and F_0 modifications. The target accented fundamental frequency contours were taken from similar utterances in readings of Cinderella in the IViE corpus [19] (with additional recordings made for Australian and New Zealand accents). The IViE corpus was chosen for this purpose as it provides accented speech from a similar distribution to the input speech (a read short story), but is a distinct set of data (thus forms a realistic scenario in which an accent transformation system is applied to unseen input data).

3.2. Presentation

It has been found that the perception of accent is not performed as a binary classification [7, 8, 9]. Rather, listeners were found to be able to perceive accent on a continuum, even when no such continuum existed in the salient accent features of natural speech. For this reason, evaluations determining the efficacy of an accent transformation approach can reasonably present participants with a range of values between two accents to select from. In these studies we have decided to present five choices, labelled *'BBC British'*, *BBC British-like'*, *'50/50 mix'*, *'X-like'*, and *'X'*, where X is the target accent. The layout for the survey is shown in Figure 1. For evaluations of RP phrases, the first target accent for that phrase given in Table 1 was taken as X. Thus for 5 of the RP utterances, the scale was from RP to one of Australian or New Zealand, and for the remaining 4 RP utterances, the scale was from RP to one of Irish or Welsh.

Participants were asked to complete a web-based form, with the utterances presented in order of appearance in the story. Participants were asked to use head phones, though it is possible

Text	Speaker	Targets
The North Wind and the Sun were disputing which was the stronger,	F4	Aus, Irish
when a traveller came along wrapped in a warm cloak.	F2	Welsh
They agreed that the one who first succeeded in making the traveller take his cloak off	M3	NZE, Aus
should be considered stronger than the other.	F1	Irish
Then the North Wind blew as hard as he could	M4	Welsh
but the more he blew, the more closely did the traveller fold his cloak around him	M5	NZE, Aus
and at last the North Wind gave up the attempt.	M2	NZE
Then the sun shone out warmly, and immediately the traveller took off his cloak.	F3	Irish
And so the North Wind was obliged to confess that the sun was the stronger of the two.	M1	Welsh

Table 1: Utterances and target accents used for the evaluation. Speakers beginning with 'M' and 'F' are male and female respectively.

Target	Pronunciation	Pron. and F_0.
NZ	3.0 (1.2)	3.3 (1.1)
Aus	3.3 (1.1)	3.6 (1.0)
Welsh	3.0 (1.3)	3.3 (1.2)
Irish	3.9 (0.8)	4.3 (0.6)

Table 2: Mean opinion scores (and standard deviations) for each target accent with modified pronunciation, and with both pronunciation and F_0 modified, as perceived by the New Zealand audience.

Target	Pronunciation	Pron. and F_0.
NZ	3.5 (1.2)	3.8 (1.0)
Aus	3.7 (0.7)	3.8 (0.8)
Welsh	2.9 (1.5)	3.0 (1.4)
Irish	3.3 (1.1)	3.9 (1.0)

Table 3: Mean opinion scores (and standard deviations) for each target accent with modified pronunciation, and with both pronunciation and F_0 modified, as perceived by the English audience.

that external speakers were used in some cases. For each input utterance, one RP version with pronunciation modified to that of Wells' analysis of RP, and F_0 modified to that of RP phrases from the IViE corpus was included. These RP utterances underwent the accent transformation process so as to ensure they were of similar sound quality to other transformed utterances. The perceived accent of the RP utterances provides a baseline against which the perceived accent of other utterances can be compared. For each target accent transformation specified in Table 1, two further versions were presented–with and without F_0 transformation. This gave a total of 21 stimuli for presentation to each participant.

3.3. Audiences

The same source and transformed utterances were presented to two audiences. The first was made up of 13 native New Zealand English speakers, living in Australia or New Zealand. The second audience was made up of 13 listeners from the north east of England, who were living in England.

4. Results

Each of the five options presented for each transformed utterance were associated with a number from 1 to 5. These then provide a familiar "mean opinion score" (MOS) of the participants' perception of the accent.

The accent of the RP versions of each utterance were given an MOS of 2.1 (where 1 is most RP-like) with a standard deviation of 0.9, by the New Zealand audience. The English audience gave an MOS of 2.2 with standard deviation 1.0. For both audiences, in all cases, the these RP utterances were scored statistically significantly lower (i.e. more RP-like) than all of the transformed versions of the utterances.

Tables 2 and 3 show the perception of accent by the New Zealand and English audiences respectively. It should be noted that the seemingly high standard deviation is largely due to the variation in relative scale between listeners. The *differences* in MOS between accents and approaches given by each listener were, however, quite consistent. So, for example, all New Zealand participants gave an MOS for transformations to an Australian accent using both pronunciation and F_0 modifications, that was close to 0.3 higher than their score for the transformation that modified pronunciation alone. Thus it is unsurprising that all of the improvements in the transformation of accent due to the inclusion of F_0 contour transformations were statistically significant by a Students' pair-wise t-test with $p < 0.01$, for the New Zealand audience. For the English audience, utterances transformed to the New Zealand and Irish accents were found to be statistically significantly more successful with the inclusion of F_0 transformations.

5. Discussion

In this discussion, we assume that the English audience is more familiar with the Welsh and Irish accents than the New Zealand audience. Similarly, it is assumed that the New Zealand audience is more familiar with the Australian and New Zealand accents. For the baseline accent–RP–we assume both audiences are reasonably familiar, though most likely the English audience somewhat more-so.

Under these assumptions, a broad comparison between the MOS of perceived accent for familiar and unfamiliar accents shows very consistent differences. Across the 8 possible comparisons between audiences (for each of the four target accents and each of the two transformation approaches), unfamiliar accents can be seen to result in MOS scores that range between 0.2 and 0.5 higher than familiar accents. More specifically, for these particular accents, and with these audiences, the MOS increases for unfamiliar accents by a mean of 0.35 with standard deviation 0.16.

Other than determining an approximate quantification of

the effect of audience familiarity on a 5-point MOS scale, it is also important to determine whether a change in audience can make significant alterations to relative successfulness. By changes in 'relative successfulness' we mean, for example, if under one audience a transformed utterance to accent A was perceived as being more successful than a transformation to accent B; is it possible that under another audience with differing familiarity, the change in perception will be sufficient such that A is no longer heard to be more successful than B?

Comparing relative success rates between accents within Tables 2 and 3, there are several possible candidates for such changes in successfulness of the transformation. For example, with an English audience transforming to a New Zealand accent was much more successful (statistically significantly, with $p < 0.01$) than when transforming to Welsh. However for a New Zealand audience, there was no significant difference between these two target accents. In a more extreme example, the Australian accent changed from being significantly more successful than Irish (for the English audience), to being significantly *less* successful (for the New Zealand audience).

A similar comparison can be made when evaluating different approaches to accent transformation. While the inclusion of F_0 modifications always resulted in a more successful change in perception for the New Zealand audience, the English audience perceived no improvement when applied to the Australian and Welsh accents. The cause of this change is not clear, however it shows that conclusions formed on the relative efficacy of differing accent transformation approaches based on evaluations with a specific audience cannot be assumed to hold across all audiences.

6. Conclusions

As expected, the familiarity of an audience with a target accent can have a significant impact on the results of perceptual evaluations of systems that transform accent. If, however, a system were evaluated on an audience unfamiliar with the target accent, the relatively consistent change in MOS scores seen in our study show that a reasonable estimate of the expected scores for an audience familiar with the accent can be made. In general, an expected drop of around 0.4 on a 5-point MOS scale is supported by the results of our parallel evaluations, when shifting to an audience more familiar with the accent.

However, we have also shown evidence for changes in comparative success rates when changing audience. Examples in which the change in audience caused an approach that was originally statistically significantly more successful at transforming accent to then be perceived as no more successful than another approach. Similarly, the relative orderings of success on specific accents was seen to change with audience familiarity. It is clear that when comparing transformation approaches, or evaluating systems targeted at a specific audience, there is no substitute for surveys presented to the target population.

7. Acknowledgements

The authors would like to thank the participants of the perceptual tests for contributing their time to help in this study. We would also like to acknowledge the Department of Computer Science at the University of Auckland, and the University of Auckland Doctoral Scholarship, that supported parts of this research completed during the first author's study toward their PhD.

8. References

[1] Street, R. L. and Giles, H., "Speech Accommodation Theory: A Social Cognitive Approach to Language and Speech Behavior", in Social Cognition and Communication, 193–226, 1982.

[2] Pucher, M., Schuchmann, G., and Fröhlich, P., "Regionalized Text-to-Speech Systems: Persona Design and Application Scenarios", in COST Action 2102 International School 2008 on Multimodal Signals: Cognitive and Algorithmic Issues, Springer LNAI 5398, 2009.

[3] Igic, A., Watson, C., Teutenberg, J., Broadbent, E., Tamagawa, R., and MacDonald, B., "Towards a Flexible Platform For Voice Accent and Expression Selection on a Healthcare Robot", in Proceedings of the 7th Australasian Language Technology Association Workshop, 2009.

[4] Böhlen, M., "Robots with Bad Accents: Living with Synthetic Speech", in Leonardo, 41(3) 209–214, 2008.

[5] Giles, H. and Powesland, P. F., "Speech Style and Social Evaluation", London, Academic Press, 1975.

[6] Grondalaers, S., van Hout, R., and Steegs, M., "Evaluating Regional Accent Variation in Standard Dutch", in the Journal of Language and Social Psychology, 29(1) 101–116, 2010.

[7] Preston, D., "Standard English Spoken Here: The Geographical Loci of Linguistic Norms", in Status and Function of Languages and Language Variants, 324–354, 1989.

[8] Plichta, B. and Rackerd, D., "Perceptions of /a/-Fronting Across Two Michigan Dialects", in NWAV 31, Stanford University, 2002.

[9] Plichta, B. and Preston, D., "The Perception of /ay/ as a North-South Stereotype in United States English", in the International Journal of Linguistics, 37:107–130, 2005.

[10] Clopper, C. and Pisoni, D., "Some Acoustic Cues for the Perceptual Categorization of American English Regional Dialects", in the Journal of Phonetics, 111–140, 2004.

[11] Fletcher, J., Grabe, E., and Warren, P., "Intonational variation in four dialects of English: the high rising tune", in Prosodic Typology. The Phonology of Intonation and Phrasing, Oxford University Press, 2004.

[12] Stylianou, Y., "Decomposition of Speech Signals into a Deterministic and Stochastic Part", in Proceedings of the International Conference on Speech and Language Processing, 2:1213–1216, 1996.

[13] Kawahara, H., Morise, M., Takahashi, Y., Nisimura, R., Irino, T. and Banno, H., "TANDEM-STRAIGHT: A Temporally Stable Power Spectral Representation for Periodic Signals and Applications to Interference-Free Spectrum, F_0 and Aperiodicity Estimation", in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 3933–3936, 2008.

[14] Teutenberg, J. and Watson, C., "Flexible and Efficient Harmonic Resynthesis by Modulated Sinusoids", in Proceedings of the European Signal Processing Conference, 2504–2508, 2009.

[15] Teutenberg, J. "On the Transformation of Accent", submitted PhD thesis, University of Auckland, 2010.

[16] Wells, J. "Accents of English", Cambridge University Press, 1982.

[17] Teutenberg, J., Watson, C. and Riddle P., "Modelling and Synthesising F_0 Contours with the Discrete Cosine Transform", in Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 3973–3976, 2008.

[18] Plante, E., Meyer, G. and Ainsworth, W., "A Pitch Extraction Reference Database", in Proceedings of Eurospeech, 837–840, 1995.

[19] Grabe, E., Post, B. and Nolan, F., "Modelling Intonational Variation in English. The IViE System", in Prosody, 51–57, 2001.