ACCENT MORPHING AS A TECHNIQUE TO IMPROVE THE INTELLIGIBILITY OF FOREIGN-ACCENTED SPEECH

Kayoko Yanagisawa and Mark Huckvale

Department of Phonetics and Linguistics, University College London, U.K.
k.yanagisawa@ucl.ac.uk and m.huckvale@ucl.ac.uk

ABSTRACT

Accent morphing aims to modify the accent of a speaker whilst maintaining speaker identity. A text-independent approach could be based on voice conversion systems which manipulate speaker identity through spectral mapping. However, it is not clear to what extent accent changes can be captured with spectral mapping alone. In this paper we implement and evaluate a text-dependent accent morphing system capable of manipulating both spectral and prosodic features. We show how accent morphing can significantly improve the intelligibility of English-accented Japanese sentences to native Japanese listeners (from 57% to 84% words correct). Analysis of the different processing conditions shows that much of the benefit of morphing comes from integrated changes to both spectrum and prosody. This suggests that text-independent morphing is unlikely to provide anything but a small increase in intelligibility.

Keywords: Morphing, foreign accent, intelligibility, speech synthesis.

1. INTRODUCTION

An accent morphing procedure would manipulate a recording of a speaker to change the speaker's accent as far as listeners were concerned, but not change those aspects of speaker identity that are unrelated to accent. If such a procedure could be developed, it would have application in the entertainment industry – for example in dubbing of foreign language films, in language learning, and in speech-to-speech translation systems. In the last case, an accent morphing technique could be applied to the output of a text-to-speech (TTS) system built for a speaker, so as to make the synthetic target language utterances – built using the source language audio data – more acceptable to listeners of the target language.

A text-independent approach to accent morphing could be based on previous work in voice conversion (e.g. [1]). Speaker A speaking accent X could be mapped to speaker B speaking accent Y using a spectral mapping trained from matched pairs of sentences. The mapping could then be applied to utterances by speaker A without knowledge of their content. This approach has a number of disadvantages: firstly the mapping may change the speech towards characteristics of speaker B which are unrelated to accent, thereby changing the speaker identity. Secondly it would not allow simultaneous modification of the pitch and timing of the utterances, so might not convey the target accent appropriately.

A text-dependent accent morphing system requires the phonetic content of the source utterance be known, either in the form of scripted or transcribed speech, or the spoken output of a translation system. If the phonetic form of the source utterance is known, then accent morphing need only be applied to those phonetic elements which are significantly different in the target accent. This will reduce the impact of the mapping on speaker identity. Knowledge of the phonetic form also means that changes to pitch and timing can also be applied.

In this study, we implement and evaluate a text-dependent accent morphing system. The source utterances are English-accented Japanese sentences, and the model utterances are native-accented Japanese versions produced by a Japanese TTS system. The morphing process involves selectively modifying aspects of the source utterances towards the model utterances. Evaluation is performed in terms of the consequences of the modifications upon the intelligibility of the morphed sentences to native Japanese listeners.

The aims of the study are to address the following questions: (i) Can accent morphing improve the intelligibility of foreign-accented speech to native listeners? (ii) What are the relative contributions of morphed pitch, timing and segmental content to any change in intelligibility? (iii) Are there any interactions between changes in segmental content and changes in pitch and timing?

2. METHOD

2.1. Source materials

The speech material consisted of 40 semantically-unpredictable Japanese sentences, each containing 4

keywords. These were adapted from [3]. Semantically unpredictable material was chosen to make the test difficult, so as to avoid ceiling effects without requiring the addition of noise.

Audio realisations of the utterances were acquired from (i) a native Japanese speaker, (ii) a Japanese TTS, and (iii) an English TTS using a custom dictionary. All versions were produced in a female voice in Standard Tokyo Japanese, at 16 kHz sampling rate. The Japanese TTS was the NeoSpeech VoiceText system using the Miyu voice. The English TTS was the AT&T Natural Voices system using the Audrey UK English voice. To make the English TTS system speak Japanese, romanised orthographic forms of the Japanese words were added to a custom dictionary. The Japanese pronunciations were entered using the best available phonetic units present in the English voice – for example, English [ʊ] for Japanese [ɯ], [ʃ] for [ɕ] and [f] for [ɸ].

A TTS system was used to produce the source utterances for a number of reasons. Firstly it was preferable to use English-accented Japanese by a speaker who knew no Japanese, so that there was no influence of level of prior language exposure. Secondly, it allows our study to be replicated using the same speaker. Thirdly, one application for accent morphing is for the manipulation of the synthetic output of a speech-to-speech translation system.

2.2. Accent Morphing

The accent morphing system takes two phonetically annotated and pitch-marked utterances and selectively transfers characteristics from the model and the source to create a new target utterance. In this experiment, phonetic labelling and pitch period marking could not be obtained from the TTS system (because we were using the SAPI interface to the systems), so phonetic labelling was performed through automatic alignment using an HMM tool (analign, in the SFS toolkit [6]). These were subsequently hand-corrected. Pitch period marking was performed using an automatic tool (SFS txanal). The best settings for this tool were optimised over the 40 sentences, but no hand correction was used.

The first stage of accent morphing was to perform pitch synchronous linear predictive coding (LPC) analysis on windows centred on each glottal impulse and of a size equal to two pitch periods. In voiceless regions, the analysis window size was chosen on the basis of a smooth interpolated pitch contour, so as to minimise large changes in window size from frame to frame through the utterance. The LP coefficients were then converted to a line spectral pair (LSP) representation, which makes the coding of the spectral envelope more amenable to interpolation across speakers. The excitation residual was extracted and stored to complement the spectral information.

The two utterances were then time-aligned so as to synchronise phonetic events. Alignment was performed using a dynamic programming procedure working from an MFCC spectral representation of the speech, but constrained by the phonetic annotations. This gave an accurate frame-by-frame alignment between source and model, even within individual segments. The morphing system then generated a new output by selecting and interpolating pitch, timing and spectral characteristics from the two input utterances.

In general, successful copying of spectral information from one speaker to another requires that the speakers have similar vocal tract sizes. However, normalisation of vocal tract sizes was considered unnecessary in this experiment, since both TTS voices appeared to have similar vocal tract sizes (in terms of their mean F4 and F5 frequencies).

2.3. Experimental Conditions

The conditions used in the experiment included the unmodified English TTS (E), Japanese TTS (J) and natural Japanese (N) versions of the sentences, together with accent-morphed variants of the English TTS. Details of the morphed conditions follow.

In the 'A' conditions, target forms with a modified spectral envelope were morphed from the Japanese TTS model and the English TTS source. The only parts of the model spectral envelope that were used were regions below 3 kHz in voiced parts. Spectral information above 3 kHz, spectral information in voiceless regions, and the excitation residual all came from the source. This was to ensure that the identity of the source speaker was affected as little as possible, as previous studies (e.g. [4]) have shown that the residual and the high-frequency spectrum contain important information about speaker identity. Time and frequency windowing was subsequently applied to give a smooth interpolation.

In the 'P' conditions, the *relative* fundamental frequency (F_0) changes for the phonetic segments were taken from the model, while mean and variance of F_0 were taken from the source. This ensured that the pitch contour was copied over but the mean F_0, important to speaker identity, was unmodified.

In the 'R' conditions, the *relative* durations of the phonetic segments in the target were taken from the model, while the overall utterance duration was taken from the source. Thus the target had the same speaking rate as the source, but modified rhythm.

As well as the individual conditions, we also looked at the combination of pitch and rhythm morphing (PR), and the combination of segment, pitch

and rhythm morphing (APR). Unfortunately, practical limitations in the size of the experiment prevented us from exploring all possible combinations. Table 1 provides a summary of the different conditions used.

Table 1: Description of each condition

E	Unmodified English TTS (source)
A	Segmental morphing alone (from J)
P	Pitch morphing alone (from J)
R	Rhythm morphing alone (from J)
PR	Pitch & Rhythm morphing (from J)
APR	Segment, Pitch & Rhythm morphing (from J)
J	Unmodified Japanese TTS (model)
N	Natural Japanese (control)

2.4. Intelligibility Test

Recordings of the 40 sentences across the 8 different conditions were randomised in a Latin-square design into 8 lists, such that each list contained 5 sentences from each condition in random order. 56 native Japanese speakers each listened to one of the lists assigned randomly, such that each list was recognised 7 times overall. Thus for each condition, word intelligibility is based on 1120 observations.

The listening experiment was conducted over the Internet, using specially-written web pages containing JavaScript functions and Java applets to control audio replay, in particular, to prevent each sentence being played more than once. Listeners typed their responses into a web form where the sentence frame was provided and only the 4 keywords needed to be completed for each sentence. Listeners were asked to input their responses using kanji and kana as appropriate, in order to disambiguate homophones which differ in pitch pattern.

A brief practice session preceded the collection of actual intelligibility data, which were collected on our web server. Responses were marked in terms of percentage keywords correct. Exact homophones with the same pitch pattern were considered as acceptable forms.

3. RESULTS

Mean intelligibility of the 8 conditions and the range are shown in Table 2 and Fig. 1 respectively. Conditions were compared in a pairwise manner using a Wilcoxon signed-rank test.

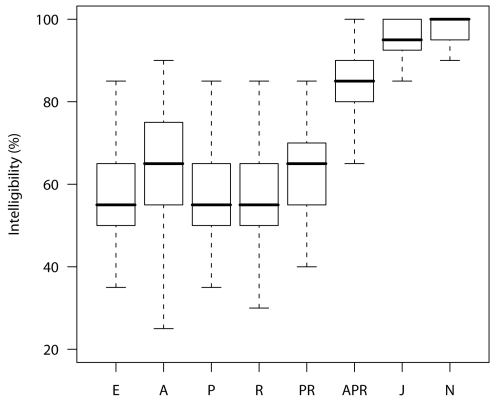
3.1. Unmodified conditions: E, J & N

As expected, the human Japanese speaker (N) gave almost perfect intelligibility scores. This control condition showed that the task and methodology

Table 2: Mean intelligibility (%, N=1120) of each condition

Cond.	Intelligibility	Cond.	Intelligibility
E	56.96%	PR	63.21%
A	64.46%	APR	84.20%
P	58.04%	J	94.91%
R	58.30%	N	95.71%

Figure 1: Range of intelligibility by condition: boxplot showing the first and third quartiles and the median.



were essentially satisfactory. The Japanese TTS system (J) also showed very good performance. A lower score would have been ideal to avoid problems with ceiling effects. Nevertheless it confirms that the Japanese TTS contains good quality segmental and suprasegmental information, adequate for use as a pronunciation target. The English TTS system speaking Japanese (E) showed considerably worse performance, as might be expected. This confirms that there is the potential for an accent morphing system to improve intelligibility.

3.2. Suprasegmental conditions: P, R & PR

Morphing the pitch of the English TTS towards the Japanese TTS (P) did not lead to a significant increase in intelligibility. This is somewhat surprising considering Japanese does use pitch information for lexical access [5]. However in this experiment, the use of sentence materials rather than isolated words may have reduced the importance of pitch information.

Morphing the rhythm towards the Japanese TTS (R) did not lead to any significant increase in intelligibility by itself either.

Interestingly, the combined manipulation of pitch and rhythm (PR) did show a small but significant increase in intelligibility ($p=0.03$) over the unmod-

ified condition (E). This could be explained by the fact that the pitch information useful for lexical access was more readily available to listeners once it was placed in the right rhythmical framework. This is supported by ideas such as segmental anchoring (e.g. [2]), which suggests that the beginning and the end of pitch movements are consistently aligned in time with identifiable landmarks. Interaction of pitch and timing has also been observed in studies such as [7].

3.3. Segmental conditions: A & APR

The modification of low-frequency spectral information in voiced regions (A) had a significant effect (p=0.007) on intelligibility over the unmodified condition (E). This change, which predominantly affects vowel realisations, clearly helps listeners identify words. However, the change is rather small. This could be due to the signal processing artefacts introduced by the incomplete source-filter separation in the analysis, leading to some vowel colour being retained in the source residual.

The combination of segmental and suprasegmental morphing causes an enormous increase in intelligibility, from 57% to 84% (E to APR), reducing the gap between condition E and condition J by two thirds. It is important to emphasise that in the APR condition, much of the source speaker characteristics was retained, as explained in 2.3.

The combination of A and PR had a considerably greater impact on intelligibility than either factor separately. This suggests that the segmental changes necessary to improve the intelligibility are different in different prosodic contexts, so that using the vowel quality of the model voice is only suitable if the prosodic environment is also correct.

Finally, the remaining gap between APR and J could have a number of causes. It could be related to the segmental information present in the voiceless regions, in the excitation residual or in the spectrum above 3 kHz. Or it may be that the morphing process itself has a deleterious effect on the signal.

4. DISCUSSION

We have described an experiment in the application of text-dependent accent morphing to improve the intelligibility of foreign-accented Japanese. The significant findings are as follows. Firstly the experiment showed that an accent morphing procedure can significantly improve intelligibility, despite possible degradation of the signal. In this experiment segmental and suprasegmental information were taken from a TTS version of the source utterance, and we targeted morphing on the low-frequency spectral envelope in voiced regions, together with pitch and rhythm. A drop of 60% in word error rate (from 43% to 16%) was achieved using this procedure.

A second outcome of the experiment is that morphing accent, pitch or rhythm individually does not have a large impact on intelligibility. This suggests that a text-independent morphing technique (applicable only to segmental changes) may be limited in its ability to improve intelligibility.

A third outcome is that the combination of segmental and suprasegmental changes has a superadditive effect over the changes individually. This interaction between the segmental and suprasegmental properties of the signal suggests that segmental changes need to be matched to the correct prosodic context to have a big impact on intelligibility. Not only does this suggest that text-independent accent morphing will be inadequate, it brings into question the ultimate performance of any speaker mapping procedure which operates solely on the basis of spectral transformation.

We hope to extend this work in two directions: firstly to investigate in more detail which specific phonetic aspects of the speech most need to be modified to improve intelligibility. The fewer elements of the source signal that we need to change, the smaller will be the impact on speaker identity. Secondly, we hope to directly compare voice transformation and accent morphing techniques on the same data, in terms of the intelligibility of the resulting speech as well as the preservation of speaker identity.

5. REFERENCES

[1] Abe, M., Shikano, K., Kuwabara, H. 1990. Cross-language voice conversion. *Proc. ICASSP-90* Albuquerque. 345–348.

[2] Arvaniti, A., Ladd, D. R., Mennen, I. 1998. Stability of tonal alignment: the case of Greek prenuclear accents. *J. Phonetics* 26(1), 3–25.

[3] Japan Electronics and Information Technology Industries Association, 2003. Speech Synthesis System Performance Evaluation Methods. JEITA IT-4001.

[4] Lin, Q., Jan, E.-E., Che, C. W., Yuk, D.-S., Flanagan, J. 1996. Selective use of the speech spectrum and a VQGMM method for speaker identification. *Proc. Int. Conf. Spoken Language Processing* Philadelphia. 2415–2418.

[5] Sekiguchi, T., Nakajima, Y. 1999. The use of lexical prosody for lexical access of the Japanese language. *J. Psycholinguisitc Research* 28(4), 439–453.

[6] Speech Filing System Tools. http://www.phon.ucl.ac.uk/resource/sfs/. visited 5-Mar-07.

[7] Ulbrich, C. 2006. Interaction of timing and pitch in cross-varietal data. *Proc. 11th Australasian International Conference on Speech Science and Technology* Auckland.