Spoken Language Conversion with Accent Morphing

Mark Huckvale & Kayoko Yanagisawa

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

Abstract

Spoken language conversion is the challenge of using synthesis systems to generate utterances in the voice of a speaker but in a language unknown to the speaker. Previous approaches have been based on voice conversion and voice adaptation technologies applied to the output of a foreign language TTS system. This inevitably reduces the quality and intelligibility of the output, since the source speaker will not be a good source of phonetic material in the new language. This article contrasts previous work with a new approach that uses two synthesis systems: one in the source speaker's voice, one in the voice of a native speaker of the target language. Audio morphing technology is then exploited to correct the foreign accent of the source speaker, while at the same time trying to maintain his or her identity. In this paper we construct a spoken language conversion system using accent morphing and evaluate its performance in terms of intelligibility. Encouraging results tell us more about the challenges of spoken language conversion.

1. Introduction

Corpus-based speech synthesis systems can now be built from the voice of any individual and are capable of producing good quality spoken realisations of any utterance in the voice of the speaker in the language of that speaker. An interesting challenge is to further develop such systems so that they can produce convincing spoken realisations of any utterance in the voice of the speaker but in a language unknown to the speaker. We call this the *spoken language conversion* problem, to distinguish it from the speech-to-speech translation problem (which aims to recognise and convert the utterance text, too) and the voice conversion problem (which aims to keep the utterance the same, but change the speaker). An earlier term was Foreign Language Synthesis [1], but this doesn't capture the idea of preserving speaker identity. Spoken language conversion systems could be used as the output component of a speech-to-speech translation system, but they could also have other applications. They might be used to produce talking phrasebooks, to dub films in a foreign language, to speak embedded foreign language phrases in a text, or to provide pronunciation targets for language learning. For the purposes of discussion, let us call the source speaker S1, the language of the source speaker L1, and the required output language L2.

What are the challenges of spoken language conversion (SLC)? Firstly the aim must be to produce L2 utterances that in the minds of impartial listeners, *could have been* produced by speaker S1. Of course the spoken language of the speaker is one of the defining characteristics of his or her identity, so we don't expect that a speaker will necessarily be recognisable when speaking L2. Anecdotal evidence is that bilingual speakers can sound like different people in their two languages. It seems likely that individuals speaking an L2 with a poor accent are more identifiable, but we don't know of evidence for this. Nevertheless, the first challenge of SLC is to preserve in L2 those aspects of the identity of the speaker that are not related to their L1 accent.

A second challenge for SLC is to generate convincing phonetic forms in L2 using knowledge only of the speaker's spoken L1. Some L1 phonetic units may make perfectly satisfactory analogues for L2 units. Most languages seem to use vowel qualities close to [i], [a] and [u] for example, and have consonants similar to [p], [t] & [k], see [2]. Other L2 units may be found by selection from a range of occasional allophonic variants exhibited by S1 in L1 – for example, a required alveolar tap [ɾ] might be found by searching through an English speaker's realisations of /t/. Some L2 units might be generated by mixing or blending sounds in L1; for example new vowel qualities might be formed by a process of interpolation between forms found in L1. Lastly, however, there may be phonetic units in L2 that have no parallel in L1 - for example retroflex stops found in Hindi - and these need to be generated by a process of extrapolation beyond forms found in L1.

A third challenge for SLC is how to deal with differences (across languages) in the phonetic interpretation of phonological units in context. The realised form of a given phonological unit will vary according to the segmental and supra-segmental environment: for example, in English, /t/ has different allophones in different syllable positions, and vowels may be reduced in different stress positions. However these very contextual variations can themselves be different across languages. Some languages do not use aspirated stops, others may or may not velarise /l/; plosives undergo lenition in some languages but not others; some languages do not exhibit vowel reduction; others may allow voiceless vowels, and so on. Languages also vary phonotactically, such that phonetic sequences found in one language might be missing from another, which in turn may lead to poorly articulated clusters. So while it may be easy to find a commonality of phonetic forms across languages in some instances, each phonetic unit also has a range of contextual variants and these variants may be different in different languages. Thus an SLC system needs to be concerned with phonetic detail at a level below that normally considered in monolingual synthesis.

A fourth challenge comes from how we ought best to evaluate the performance of SLC systems. The recent tendency for the evaluation of monolingual synthesis systems has been the use of a mean opinion score (MOS), using a rating scale from 1-5. Such an approach is not without problems when applied to SLC. If we used MOS to evaluate SLC systems we would, of course, need to use native listeners of L2 for the rating. However SLC systems also need to be evaluated in terms of how well speaker identity is preserved,

and this raises issues about how well individuals can be recognised when speaking another language anyway. In addition, if we seek to compare different SLC systems, it may be hard to disentangle the perceptual consequences of processing artefacts from the assessment of speaker similarity. Listeners may be more critical of a clean and precise synthesis of S1 in L2 that does not express S1's identity exactly, than a noisy, messy synthesis where identity is less easy to establish anyway. Finally, MOS experiments require a large pool of listeners, which make them expensive to perform. They can also be insensitive to small variations in system performance [3].

In this paper we will review previous technological approaches to the spoken language conversion problem. We will try to highlight what we see as their limitations. We then introduce a new approach based on accent morphing – a process that involves interpolation between two versions of a spoken utterance. We demonstrate the potential of accent morphing within the context of spoken language conversion by showing how well it improves the intelligibility of foreign-accented TTS synthesis to native listeners. We conclude by drawing some implications for the construction of future spoken language conversion systems.

2. Previous approaches to Spoken Language Conversion

Any synthesis system that can be controlled at a phonetic level can be made to simulate a foreign language simply by selection of appropriate units from the L1 inventory. We don't consider such approaches here since they will have severe foreign accents, although they might function as control conditions in SLC experiments. Perhaps the first approach to SLC that went beyond phonological selection from L1 was Campbell's foreign language synthesis system [1]. This system was based on the CHATR corpus-based synthesis system, but modifications were made at the level of unit-selection so as to choose corpus units for synthesis (from L1) that were best suited to implementing the required phonetic forms in L2. In conventional unit-selection, candidate units are selected on the basis of a phonological match to the target utterance. It is assumed that the phonetic detail in the selected speech signal sections is appropriate because of the match in phonological labels. For foreign-language synthesis, we can map the phonological labels, but this does not guarantee the appropriateness of the phonetic detail. Campbell's approach was to use a phonetic target for unit selection based on acoustic analysis of a synthesized native version of the utterance. Unit-selection then becomes a process to choose among phonetic units rather than phonological units. In terms of how well Campbell's system meets the challenges of SLC, we note that S1's voice is used in an unmodified form, and so in one sense S1's identity is maximally preserved. However since the process only selects from S1's available units, it does not address the problem of L2 units which are poorly realised or missing in the source system. While the acoustic matching to L1 might provide some appropriate contextual variants, it can't deal with contexts or variants that are missing in L1. Evaluation of the system was very limited, and performed only in terms of MOS on isolated words with no control condition.

The advent of speech-to-speech translation systems in the 1990s encouraged the development of speaker-adaptable text-to-speech systems: synthesis systems which were implemented in language L2 using some different speaker S2, but which could be modified to sound like S1. The dominant technique for this adaptation was then, and remains today, *voice conversion*. In voice conversion, an utterance is modified by some signal processing techniques to change the identity of the speaker, but to leave the linguistic content of the utterance unchanged. A number of voice conversion approaches have been proposed, e.g. [4,5,6]. All these techniques have at their heart a statistical model which maps spectral details across two speakers. An utterance spoken by speaker S2 is broken down into spectral vectors, then each of these is substituted by vectors estimated as representative of speaker S1 and the utterance resynthesised. The training of the mapping from S2 to S1 is performed by aligning equivalent speech signals in training data produced by S2 and S1. Gaussian mixture modelling of LPC-derived spectral envelopes is a common technique.

Voice conversion as described above is really only suited for mapping between speakers that speak the same language – this is because the mapping is learned from a training corpus of matched signals, and the matching relies on a phonetic equivalence of the signals. Attempts have been made to adapt voice conversion across languages, for example [7,8,9]. Mashimo [8] used a trick based on a bilingual speaker S2 who could speak both L1 and L2. A text-to-speech system was implemented in S2's voice in language L2, but then the voice conversion mapping was learned between S1 and S2 speaking L1. This allowed for the mapping to be learned from matched sentences spoken by both S2 and S1. Sündermann et al [9] adapted the idea so that the matched sentences in L1 were generated by unit-selection from a corpus of speaker S2 speaking L2. To understand the performance of these cross-language voice conversion systems, we need to understand more about how phonetic equivalence across languages is established. If for example, the mapping is learned from materials that are the same only in terms of phonological transcription using a phoneme-level association across languages, then it is likely that this mapping will fail to accommodate differences in phonetic detail. If, for example, voice conversion changed a native [ɾ] to a foreign [ɹ] to implement /r/, then intelligibility of the L2 utterances may suffer. This is just one example of a general issue about context sensitivity in cross-language voice conversion. Since the whole approach is based on estimating a single best spectral slice in S1 for a spectral slice found in S2, then there is no mechanism for the mapping to be made sensitive to the phonetic, phonological or prosodic context of the utterance. The 'best' mapped spectral slice may be different in different contexts: whether this is part of an /l/ or an /r/, whether it is in a stressed syllable or an unstressed one, whether it is phrase final or phrase initial, and so on. Evaluation of Sündermann's system indeed shows that MOS ratings after conversion are much lower than before. The process of cross-language voice conversion reduces the rating of the synthetic speech from 4.7 to 3.5. Worse, this reduction in quality does not seem to be matched by a large increase in the rating of S1 speaker similarity, here the MOS only increased from 1.6 to 2.0 after voice conversion. This may be because current voice conversion technology finds it easier to map overall spectral envelopes rather than details of the speaker's source signal [10].

Recently a third technology has been developed that could be capable of spoken language conversion. Latorre et al [11] describes an HMM synthesis system which is trained using multiple voices, and adapted using a single target voice. If such a synthesis system were trained with multiple languages, using an extended phone set to achieve a consistent labelling, then the approach could be used to generate a number of languages in one new target voice. The key difference to voice conversion is that adaptation is performed at the level of phones rather than at the level of spectral slices. This provides a level of context sensitivity, whereby the same spectral detail in two different phones might be mapped to different values. To perform the adaptation, a set of phonologically labelled utterances from S1 in L1 are fed into the system to adapt all the phone models even though only some of those phones in only some contexts will be present in the adaptation utterances. It seems that within the system, phones (across all languages) are clustered into groups, and a linear transformation of spectral means are applied to all units within a cluster, estimated from the adaptation material. It is not clear how this process affects the foreign language phones not present in L1, and the impact these have on intelligibility. In terms of preserving the identity of S1, Latorre's system is somewhat hampered by the relatively poor voice quality of HMM synthesis compared to corpus synthesis. However, HMM synthesis could use samples of S1's LPC residual to excite each phone model, and this could improve the identifiability of the speaker. Once again, the use of "equivalent" phonetic forms across languages, even when their precise realisation will be different in context, means that Latorre's system will also replace correct L2 forms with L1 approximations, leading to a reduction in intelligibility. Consider an L2 which uses [t^h] in one environment and [t] in another, if the adaptation process replaced both with a particular implementation of /t/ in L1, then the adapted speech will end up with incorrect detail. This type of effect could explain the reduction in the MOS of the L2 speech after adaptation (from 4.3 to 3.8), even when the MOS rating of identity improves (from 2.6 to 3.1).

In this section we have seen three approaches to spoken language conversion. We suggest that all have some weaknesses, many related to the use of an overly simplistic model of the phonetic relationships between languages. A table of phonological equivalences is not going to be good enough when the realisations of those units depends on the contexts in which they occur and in which language they are produced. The aim of our research is to explore these mismatches in more detail, and to that end we have developed another approach to spoken language conversion which provides more control over the phonetic mapping between L1 and L2.

3. Accent Morphing

The long term objectives of our research are to give a quantitative account of the differences between accents, both regional accents and foreign language accents. Spoken language conversion is a convenient testing ground for ideas about what aspects of accent are most salient to listeners. For any language pair, we can use the technology to generate and compare arbitrary utterances, then we can evaluate the consequences of differences in phonetic detail between them. Particularly we want to study how differences in phonological inventory and phonological interpretation across languages have an impact on the intelligibility and acceptability of a speaker S1 producing L2. To do this we needed a model of L2, a model of speaker S1 and the ability to control the phonetic composition of new utterances.

Our first insight was that the best knowledge we have for how to produce an utterance in L2, complete with all appropriate contextual variation, is through the use of a synthesis system in L2. So we use an L2 text-to-speech system as a knowledge source for how to speak L2, just like Campbell [1]. Similarly, the best knowledge we have about speaker S1, complete with how they produce different phonetic forms in different contexts, is through a synthesis system built in the voice of speaker S1. Inevitably this latter system will be in language L1, since we assume that speaker S1 does not speak L2.

Using our two text-to-speech systems, we can now generate a foreign-accented version of some target utterance U1 using system S1L1, and we can generate a native-accented version of the utterance U2 using system S2L2. If we could establish which aspects of U1 are inappropriate or inadequate, say by comparing it to U2, we can perform a signal processing transformation on only those aspects of U1 which need to be changed. The advantage of this is that U1 remains in the voice of speaker S1, and those aspects that are satisfactory are unmodified in the procedure. We call this technique *accent morphing*, because it takes as input two versions of the same utterance and generates a third version which borrows speaker information from one and accent information from the other. In other words, we implement a spoken language conversion system by generating the target L2 utterance using S1's voice, and then "patching up" the inevitable foreign accent in such a way as to minimise the impact on his or her identity.

How can we establish which aspects of U1 need to be changed? We have two sources of information: general information about the phonetics and phonology of the two languages, and specific information about the spectral qualities used in the utterances U1 and U2. We might, for example, simply identify particular phones which are likely to be problematic. On the other hand we might be able to use knowledge of accent variability and human perception to judge whether the existing implementation of a phone in U1 is within an acceptable range. The work done by Huckvale on the ACCDIST metric for comparing accents across speakers [12] might be used to establish which segmental qualities are furthest from the norm for the target accent.

How can we perform the signal modifications appropriate for this utterance? We might do this by "borrowing" temporal and spectral information from U2 and blending it with U1. For example, we might match vocal tract sizes across S1 and S2, so that we can predict target spectral envelopes for some phone in L2 in this context. A number of possible technical approaches could be taken to perform accent morphing. Techniques based on LP analysis and residual excitation seem practical [13]. We describe one particular implementation in the next section, although we are sure that better methods will be developed in the future. The concept presented here is not specific to some particular form of signal processing. However the spectral manipulation is performed, it only needs to be applied in some phonetic contexts and can be made sensitive to the requirement to preserve the identity of speaker S1.

How does this approach meet the challenges of SLC? Firstly it aims to re-use the speech of S1 in all places where it is satisfactory, this may mean re-use of the source signal, or of some whole segments or even of some frequency regions within segments. Information about phonetic units missing in L1 can be borrowed from U2, and furthermore, these will have appropriate contextual forms for L2. Lastly, we know that foreign accents are less intelligible to native listeners, therefore we can evaluate success by measuring the increase in intelligibility brought on by accent morphing. The next section evaluates one implementation of the idea.

4. Intelligibility Experiment

4.1. Aims

This experiment was designed to see if it is possible to implement an accent morphing system as part of a spoken language conversion application, and to assess the intelligibility of its output. Specifically, we addressed the following questions: (i) Can accent morphing improve the intelligibility of foreign-accented TTS output 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? This experiment did not address the impact of accent morphing on speaker identity, which is left for a further study. However we have tried as far as possible to minimise the impact of the processing on identity.

4.2. Source materials

The speech material consisted of 40 semantically unpredictable Japanese sentences, each containing 4 key words. These were adapted from [14]. Semantically unpredictable material was chosen to make the test difficult, so as to avoid ceiling effects in intelligibility scores, 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.

4.3. Accent Morphing

The accent morphing system takes two phonetically annotated and pitch-marked versions of an utterance: one from the source speaker and one from the model speaker. These are analysed and aligned and then used to generate a new target version of the utterance by selecting and combining characteristics from them. In this experiment, phonetic labelling and pitch period marking could not be obtained from the TTS systems (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 [15]). 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.

Analysis consisted of pitch synchronous linear predictive coding (LPC) 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 provide continuity in analysis window size from frame to frame through the utterance. The LPC coefficients were then converted to a line spectral pair (LSP) representation, to make the coding of the spectral envelope more amenable to interpolation across speakers. The excitation residual was extracted from the source speaker for each separate glottal cycle and stored to complement the spectral information.

Alignment of the utterances 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 cycle-by-cycle alignment between source and model speaker versions of the utterance, even within individual segments.

Morphing was then performed by generating the target utterance one glottal cycle at a time by selecting and interpolating pitch, timing and spectral characteristics from the set of aligned glottal cycle pairs. For some output time t, the corresponding source time is found from the required target timing. Similarly the synthesis window offset from the previous output cycle is found from the required target pitch. The required spectral envelope is found by interpolation of the envelopes of the matched cycles, while the required residual is just copied from the source speaker. Resynthesis from the interpolated LSP parameters and residual is then performed by overlap-add. 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 size was considered unnecessary in this experiment, since both TTS voices appeared to have similar vocal tract sizes (assessed in terms of their mean F4 and F5 frequencies).

4.4. Experimental Conditions

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)

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 as model speaker and the English TTS as source speaker. The only parts of the model spectral envelope that were used were regions below 3.5kHz in voiced parts of the sentence. Spectral information above 3.5kHz, spectral information in

voiceless regions, and the excitation residual all came from the source speaker. This was to preserve the identity of the source speaker as much as possible, consistent with modifying phonetic quality towards the model. Previous studies (e.g. [16]) have shown that the residual and the high frequency spectrum contain important information about speaker identity. To limit artefacts arising from the switching of speaker data across and within frames, windowing was applied. Time windowing occurred across a single glottal cycle at the start and end of each voiced section, while frequency windowing extended from 3000 to 4000Hz, both using a linear interpolation.

In the 'P' conditions, the relative fundamental frequency (F0) changes for the phonetic segments were taken from the model speaker, while mean and variance of F0 were taken from the source. This ensured that the pitch contour was copied over but that the mean F0, 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.
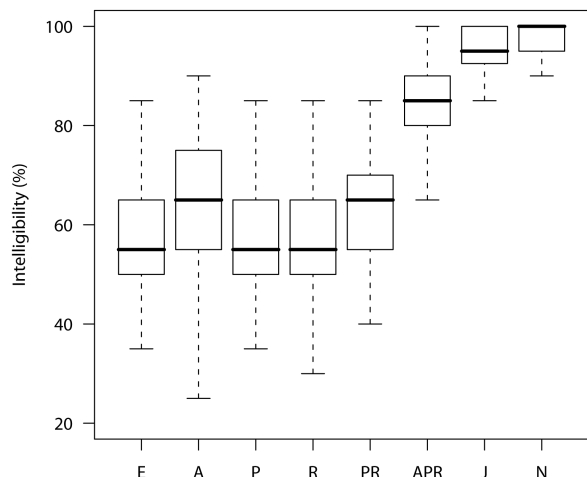
4.5. Intelligibility Test

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 prevent each sentence being played more than once. Listeners typed their responses into a web form where the sentence frame was provided and only 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.

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

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: Word intelligibility by condition

4.6. Results

The distribution of intelligibility scores across conditions is shown in Fig 1, and the means are summarised in Table 2. Conditions were compared in a pairwise manner using a Wilcoxon signed-rank test.

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 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.

Suprasegmental conditions: P, R & PR

Morphing just the pitch of the English TTS towards the Japanese TTS (P) did not trigger a significant increase in intelligibility. This is somewhat surprising considering Japanese does use pitch information for lexical access [17]. However in this experiment, the use of sentence materials rather than isolated words may have reduced the importance of pitch information. Morphing just the rhythm of the English TTS towards the Japanese TTS (R) also did not produce a significant increase in intelligibility. However the combined manipulation of pitch and rhythm (PR) did show a small but significant increase in intelligibility (p=0.03) over the unmodified condition (E). These facts might be explained if pitch information useful for lexical access was more readily available to listeners once it was placed in the right rhythmical framework. The interaction of pitch and timing like this has also been observed in studies such as [18].

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 caused by segmental quality change alone is rather small. One explanation for this might be due to morphing artefacts. For example an incomplete source-filter separation in the analysis could lead to some vowel colour being retained in the source residual.

The combination of segmental and suprasegmental morphing caused a large increase in intelligibility, from 57% to 84% (E to APR), reducing the gap between condition E and condition J by two thirds. Perhaps it is important to emphasise here that in the APR condition, much of the source speaker characteristics were still retained, as explained in 4.4. 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 segmental quality of the model voice is only suitable if the prosodic environment is also correct. Finally, the remaining gap between conditions 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.

4.7. Discussion

We have described an experiment in the application of accent morphing to improve the intelligibility of foreign-accented Japanese TTS. The significant findings are as follows. Firstly the experiment showed that an accent morphing procedure can significantly improve intelligibility, despite any degradation in signal quality that may have been caused by signal processing. In this experiment segmental and suprasegmental information were taken from a Japanese 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 finding of the experiment is that morphing pitch or rhythm or segmental quality separately has surprisingly little effect on intelligibility. The lower intelligibility of the English TTS system speaking Japanese is not due to just one of these factors.

A third finding is that the combination of segmental and suprasegmental changes has a superadditive effect on intelligibility over the changes individually. This clear demonstration of an interaction between the segmental and suprasegmental properties of the signal is further evidence that phonetic differences between languages are contextually conditioned. It is only when the Japanese segmental forms are used in the right Japanese prosodic contexts that they significantly improve intelligibility.

5. CONCLUSIONS

In this paper, we have introduced a new approach to building a spoken language conversion system: a TTS system in L1 is used to produce L2 then the worst aspects of its foreign accent are corrected using accent morphing. The experiment we have presented did not evaluate a complete SLC system but concentrated on how phonetic differences between languages can have an impact on intelligibility. We

have shown that the technique can produce highly intelligible Japanese utterances from an English TTS system. Detailed results also show that there are segmental and suprasegmental differences and segmental-suprasegmental interactions which need to be accommodated in a spoken language conversion system. For this particular language pair, we find that segmental quality changes alone do not have a large benefit. This suggests that the spectral mapping of the kind employed in voice conversion systems - which is applied separately from a change in prosody - may limit their ability to improve intelligibility. We have also shown that phonetic details need to be matched to the prosodic context – only when the two are in step do we see a significant improvement in the output. This suggests that a speaker adaptable TTS system that operates across languages may need to condition segmental adaptations on the prosodic context in which they occur.

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.

6. REFERENCES

[1] Campbell, N., "Foreign Language Speech Synthesis", *3rd Speech Synthesis Workshop*, Australia,, 1998, 177-181.

[2] Ladefoged, P., Maddieson, I., "Sounds of the world's languages", *Blackwell Press*, 1996.

[3] Vazquez-Alvarez, Y., Huckvale, M., "The Reliability of the ITU-P.85 Standard for the Evaluation of Text-to-Speech Systems", *Proc. ICSLP-2002*, Denver, 2002.

[4] Abe, M., Nakamura, S., Shikano, K., Kuwabara, H., "Voice conversion through vector quantisation", *J. Acoust. Soc. Japan*, 11(2), 1990, 71-76.

[5] Stylianou, Y., Cappé, O., Moulines, E., "Statistical Methods for Voice Quality Transformation", *Proc. EuroSpeech 1995*, Madrid, Spain, 1995.

[6] Toda, T., Lu, J., Saruwatari, H., Shikano, K., "Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum", *Proc ICSLP-2000*, Beijing, 2000, 279-282.

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

[8] Mashimo, M., Toda, T., Kawanami, H., Kashioka, H., Shikano, K., Campbell, N., "Evaluation of Cross-Language Voice Conversion using Bilingual and Non-Bilingual Databases", *Proc EuroSpeech-2001*, Aalborg, 2001, 361-364.

[9] Sündermann, D., Höge, H., Bonafonte, A., Ney, H., Hirschberg, J., "Text-Independent Cross-Language voice Conversion", *Proc. ICSLP 2006*, Pittsburgh, USA, 2006.

[10] Sündermann, D., Höge, H., Bonafonte, A., Ney, H., and Black, A., "Residual Prediction Based on Unit Selection", *Proc. of 9th IEEE Automatic Speech Recognition and Understanding Workshop*, San Juan, Puerto Rico, 2005.

[11] Latorre, J., Iwano, K., Furui, S., "New approach to the polyglot speech generation by means of an HMM-based

speaker adaptable synthesizer", *Speech Communication*, 48, 1227-1242, 2006.

[12] Huckvale, M., "ACCDIST: a metric for comparing speakers' accents", *Proc. ICSLP-2004*, Jeju, Korea, 2004.

[13] Ye, H., & Young, S. "High quality voice morphing". *Proc. ICASSP 2004*. Canada, 2004, Vol 1, 9-12.

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

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

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

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

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