A Phonetic Alternative to Cross-language Voice Conversion in a Text-dependent Context: Evaluation of Speaker Identity

Kayoko Yanagisawa & Mark Huckvale

Department of Speech, Hearing and Phonetic Sciences, UCL, London, U.K.
kayoko.yanagisawa@uclmail.net, m.huckvale@ucl.ac.uk

Abstract

Spoken language conversion (SLC) aims to generate utterances in the voice of a speaker but in a language unknown to them, using speech synthesis systems and speech processing techniques. Previous approaches to SLC have been based on cross-language voice conversion (VC), which has underlying assumptions that ignore phonetic and phonological differences between languages, leading to a reduction in intelligibility of the output. Accent morphing (AM) was proposed as an alternative approach, and its intelligibility performance was investigated in a previous study. AM attempts to preserve the voice characteristics of the target speaker whilst modifying their accent, using phonetic knowledge obtained from a native speaker of the target language. This paper examines AM and VC in terms of how similar the output sounds like the target speaker. AM achieved similarity ratings at least equivalent to VC, but the study highlighted various difficulties in evaluating speaker identity in a SLC context.

Index Terms: accent morphing, voice conversion, spoken language conversion, speaker identity, cross-language

1. Introduction

Spoken language conversion (SLC) is the challenge of using speech synthesis systems and speech processing techniques to generate utterances in the voice of a target speaker, but in a language that they do not actually speak (L2). It has applications in speech-to-speech translation, dubbing of foreign language films and in foreign language learning. In a text-dependent context — for example, in the speaker-adaptation of a text-to-speech (TTS) system in another language — two approaches are conceivable: one which takes a TTS voice in L2 and converts the speaker characteristics towards the target speaker, and another, which takes an L1 TTS voice speaking L2 with an L1 accent, and corrects the accent using the L2 TTS. They both make use of a speaker of L2 (resource speaker), but in different ways.

The mainstream approach thus far has been the first approach, namely cross-language voice conversion (VC), which is based on statistical models that map spectral details across speakers to convert the speaker characteristics. However, it has been pointed out that they often ignore phonetic and phonological differences between languages [1].

The problem with VC in a SLC context is that the target speaker, whose native language is L1, does not speak L2, and thus parallel utterances are not readily available for training the mapping. Mashimo et al. [2] proposed an approach using a bilingual resource speaker who could speak L1 and L2, and trained a conversion function between the resource speaker and the target speaker based on parallel utterances in L1 by the two speakers. They then applied the spectral mapping data to the resource speaker speaking L2, to make it sound like the target speaker. Thus the training and the conversion were carried out in different languages. Since the phoneme inventory of two different languages are never the same, not all spectral characteristics in L2 may be captured in L1. Sündermann et al. [3] described a technique to create pseudo-parallel utterances in L2 between the resource speaker speaking L2 and the target speaker speaking L1, by selecting the best matching spectral slice from the latter for each spectral slice in the former. This made it possible to carry out the training and conversion in the same language (L2), but the conversion function was trained between L2 sounds on the one hand and acoustically similar L1 sounds on the other, leading to assumptions of equivalence between sounds of the two languages. Thus the accuracy of the mapping function would depend on the match between the resource and target frames. In a previous study [1], we found that cross-language VC reduces the intelligibility of the output, due to the phonetic equivalence assumptions. From the speaker identity point of view, since the mapping is based solely on the acoustic similarity of the frames, the more similar corresponding resource and target frames are, the less speaker-dependent information can be trained from the training utterances. In the extreme case, the conversion would output utterances identical to the input resource speaker.

Subsequently, we proposed an alternative approach, accent morphing (AM). AM attempts to preserve the voice characteristics of the target speaker, whilst modifying their accent using phonetic knowledge obtained from an L2 TTS system, the resource speaker. It takes parallel utterances, one by the target speaker speaking L2 with an L1 accent, and the other by the L2 resource speaker. It then uses audio morphing technology to change those aspects of the target speaker speech signal which are related to accent, to modify the accent characteristics towards the resource speaker's, whilst preserving as much of target speaker characteristics as possible. Essentially, it exploits the fact that some aspects of the speech signal are more strongly correlated with accent than with speaker identity. It was shown in [4] that AM is capable of improving the intelligibility of foreign-accented speech.

The objectives of this study were to compare the performance of AM and VC in terms of speaker identity, and to tackle the challenges in the evaluation of SLC systems, which necessarily involve multiple languages and accents.

2. Method

An AM system and a VC system were implemented to generate utterances in Japanese as if spoken by a monolingual English target speaker. Except where explicitly mentioned, all signal processing was carried out using the SFS toolkit [5].

2.1. Speakers

The target speaker (TS) was an English TTS system (AT&T Natural Voices system with the female Audrey UK English voice). Both AM and VC made use of a Japanese resource speaker (RS), either as a model to obtain the accent information from (in the case of AM), or as a source to convert the speaker identity from (in the case of VC). This was a Japanese TTS system (NeoSpeech VoiceText system with the female Miyu voice).

2.2. Evaluation Materials

The evaluation material consisted of 18 Japanese sentences, randomly selected from the ATR 503 phonetically balanced sentences [6] so that the average durations for each sentence produced by TS and RS were both about 8 seconds. Audio realisations of the utterances were acquired from the English TTS (TS) and the Japanese TTS (RS). All versions were produced at 16 kHz sampling rate.

2.3. Accent Morphing

The AM system started from the English-accented TS version of the sentences, and attempted to correct the accent whilst preserving the speaker characteristics.

2.3.1. Preparation

To make the English TTS 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 English, whereby the mapping was established by rules. Phones with the same IPA representation were mapped to each other. Where no equivalent phone was found, the phone with the most articulatory features in common was chosen, and the perceptual similarity of the corresponding phones was checked by a native Japanese speaker (the first author). Such mapping by rule has been found to be more reliable than mapping based on acoustic distances, in other cross-language work requiring phone mapping e.g. [7].

Phonetic labelling of the evaluation sentences by TS and RS was performed through automatic alignment using an HMM tool in SFS, followed by hand-correction. Pitch periods were then marked automatically using Praat [8]. The utterances were submitted to a 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 LPC 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 from TS for each separate glottal cycle and stored to complement the spectral information.

2.3.2. Morphing and resynthesis

The paired TS and RS utterances were then time-aligned so as to synchronise phonetic events. Alignment was performed using a dynamic programming procedure working from a mel-frequency cepstral coefficient (MFCC) spectral representation of the speech, but constrained by the phonetic annotations. This gave an accurate frame-by-frame alignment between the resource and target utterances, even within individual segments.

In order to simulate a normalisation for vocal tract length, which contributes to speaker identity, the RS LSP parameters were scaled towards those of TS before morphing. The mean of F4 and F5 for TS was 95.9 % that for RS. Therefore RS LSPs were scaled correspondingly.

Morphing was 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. Various AM conditions were tested, but only one implementation will be presented here. In this implementation, the following information was copied from RS:

- Relative f0 changes (pitch contour)
- Relative durations of the phonetic segments
- Spectral envelope below 2.5 kHz[1] in voiced regions.

All other aspects of the speech signal, including the mean f0, overall utterance duration, residual, spectral envelope above 2.5 kHz and the entire spectral envelope in voiceless regions was taken from TS, in order to preserve speaker identity. Resynthesis from the interpolated LSP parameters and the residual was then performed by overlap-add.

2.4. Voice Conversion

The VC system started from the RS version of the sentences and attempted to convert the speaker characteristics towards TS. A standard VC procedure, based on a Gaussian mixture model (GMM) and a set of linear transforms was used.

2.4.1. Training data

The training material consisted of 100 paired English sentences from the two speakers, each containing 4 words. They were designed to provide complete coverage of those Japanese phonemes which have English "equivalents". To achieve this, real English words were chosen from the BEEP dictionary [9] to meet two constraints: (i) that the English phonemes used have Japanese equivalents, and (ii) that the English words did not violate the phonotactics of Japanese. This resulted in a list of words such as "adjourning, marginally, leukaemia, knocking" which could be transliterated into Japanese and spoken by RS, as well as read in English by TS. The phone equivalence which was assumed in transliterating the English words into Japanese was established by rules, as described in Section 2.3.1.

2.4.2. Preparation and training

The training and conversion of linear transformation parameters followed the model proposed by [10]. Firstly, pitch-synchronous LSP analysis was applied, with pitch period marks automatically obtained by Praat. The training sentence pairs were then aligned using a dynamic programming algorithm based on the Euclidean distance based on MFCC representation. A GMM with 16 mixtures and diagonal covariance was trained for RS alone, using a spectral representation comprising 18 LSPs, and with one estimated linear transform per mixture. Only 16 mixtures could be trained from the limited data.

2.4.3. Conversion

The 18 Japanese evaluation sentences acquired from RS were then converted towards TS. The final transformation for a resource vector was estimated from a weighted sum of the set of transforms, according to the mixture probabilities.

[1] More precisely, the transition took place over a 1 kHz region centred on 2.5 kHz.

The f0 trajectory of the RS utterance was modified by a factor of 0.83, which was the ratio of the mean f0 of RS to TS in the training data. Perceptual filtering was applied to the transformed spectral envelope, following [11], to attenuate the noise in the spectral valleys and to therefore make the formant peaks clearer and the speech more intelligible. The LSPs were thus mapped and the RS's residual was then put through the new LSP filter to generate the output signal.

3. Evaluation

3.1. ABX

3.1.1. Task

Forty-seven Japanese listeners, who had no known hearing impairment, and were not familiar with TS or RS, took part in an ABX task. They were presented with triplets of utterances X, A and B, where X was the unmodified English of TS and A and B were 2 versions of the converted utterance (AM or VC) in Japanese. Listeners were asked to judge whether X sounded closer to A or B in terms of speaker identity. The order of presentation (X – AB or X – BA) was randomised and listeners listened to each pair in both order 3 times. Each sentence of the triplet lasted about 8 seconds, and listeners could listen to them as many times as they liked.

3.1.2. Results

If listeners always made reliable judgements, their preference should always be the same for identical triplets. Therefore, the self-agreement rate, showing the consistency of the listener's judgement on these identical triplets, was calculated for each listener. A reliable judgement would need to have a self-agreement rate beyond the 50 % chance level. Cohen's Kappa coefficient [12] was calculated to factor out the proportion of agreement that could be expected by chance alone. Following [13], a κ value above 0.41was considered to be in "moderate agreement" and thus reliable. 13 listeners out of 47 met this criterion, and these were used for subsequent analyses. The results showed that, when presented with condition VC and AM in either order, listeners judged AM to be more similar to TS in 65.4 % of the cases. A chi-square test confirmed this preference of AM over VC to be significant ($p = .01$). The two conditions were found to be comparable in terms of intelligibility in a separate experiment (not presented here).

However, there are a number of problems with this evaluation method. Firstly, the results indicate that AM is better than VC in generating output that sounds like TS, but no indication is given as to the similarity in absolute terms. Secondly, the triplets consisted of a reference sentence X in English followed by Japanese sentences in AM and VC. Due to the way in which the signals are manipulated, AM output retains some degree of English accent. This may have worked in favour of AM when listeners made their judgements, even though they were given explicit instructions to ignore any accent differences.

3.2. Rating

3.2.1. Task

To overcome these problems, another evaluation task was designed. In this paired similarity ratings task, listeners rated the perceived similarity of the SLC outputs to RS as well as to TS on a scale of 1 to 7, where 1 = "completely different" and 7 = "very similar". Each condition was presented in pairs with TS and RS to 45 native listeners of Japanese. Listeners listened to each pair 3 times, each time in a different sentence. The order of presentation of pairs was randomised in a Latin-square design, as well as the combination of the sentences and conditions. For each pair, the reference sentence was presented before the evaluation sentence. The reference sentence was either TS speaking English or RS speaking Japanese.

The evaluation sentences were in Japanese, while the reference sentences were in Japanese for RS but inevitably in English for TS, giving rise to a problem similar to the accent problem that came to light in the ABX experiment. It is easier to compare utterances in the same language than those in different languages, since listeners will also have access to linguistic cues to speaker identity such as accent, in addition to various non-linguistic attributes. This would result in a bias of similarity towards RS. In order for speaker identity to be evaluated independently from intelligibility, evaluation sentences were temporally reversed. Temporal reversal alters the speech signal in such a way that linguistic information is harder to extract, but leaves speaker-dependent attributes such as the long-term average spectrum, vocal quality, speaking rate and f0 mean and range unaltered. As a consequence, it is still possible to recognise the speaker through reversed speech, as found in previous studies such as [14, 15].

3.2.2. Results

As shown in Table 1, the VC output had a mean TS similarity rating of 4.03, whereas AM was rated 4.31. A one-way repeated measures ANOVA followed by Bonferroni *post hoc* tests showed that, contrary to the ABX test results, there was no significant difference between VC and AM in terms of TS similarity.

Table 1: *Mean and standard deviation of ratings of similarity to the target speaker (N = 135 per condition).*

Condition	E	J	VC	AM
Mean	6.23	3.40	4.03	4.31
SD	0.77	1.47	1.18	1.15

Condition E was TS speaking Japanese with an English accent, and condition J was RS speaking Japanese natively. For VC, the *post hoc* tests showed that there was a significant difference between E and VC ($p < .001$) but not between J and VC ($p = .29$). This means that VC was not able to generate an output sounding as TS-like as TS herself (condition E), and that it may not be worth applying VC to RS utterances (condition J), since the conversion reduces the intelligibility whilst not improving on the TS similarity. For AM, it was found that there was a significant difference between E and AM ($p < .001$) as well as between J and AM ($p = .005$). AM reduced the TS similarity of the output, but the output was not as dissimilar to TS as RS was.

In terms of similarity to RS (as shown on the x-axis in Fig. 1), both VC and AM were successful in making the output sound less like RS than RS herself (J), but there was no significant difference between VC and AM ($p = 1.00$).

4. Discussion

The rating experiment showed that AM and VC, which operate in opposite directions, achieved a similar level of performance

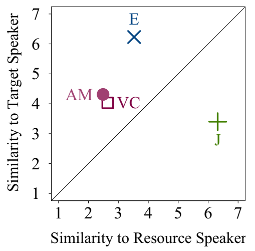

Figure 1: *The centroids of mean ratings of similarity to the target speaker and the resource speaker. The diagonal is the line for which any point above it is more similar to TS than to RS.*

in terms of similarity to TS and RS. Fig. 1 shows both points above the diagonal, indicating that they were both successful in generating output sounding more like TS than RS. The RS similarity ratings for these points are considerably lower than for condition J, but in terms of TS similarity, they are only about a third of the way between J and E, and far from sounding like TS. This may be a reflection of the non-linear way in which humans perceive speaker similarity, in that there is only a narrow margin in which listeners perceive two speakers to be similar, and once outside this margin, similarity rating drops immediately. Thus it is difficult to generate an output that sounds like a particular speaker, although it is easy to make it sound *un*like a particular speaker.

Another outcome of this study is the variability with which listeners perceive speaker similarity, as shown by the dispersion of the points on Fig. 2. This is partly due to the fact that reversed speech was used, but listeners also relied on different cues in making similarity judgements. In an informal questionnaire asking what cues they had used, it was revealed that some listeners still relied on subtle accent differences they had detected in the reversed speech, despite being asked to ignore any such differences.

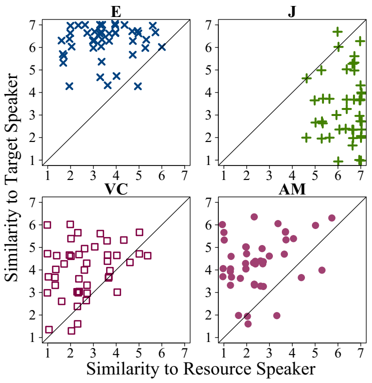

Figure 2: *Scatterplot of rating of similarity to the target speaker against similarity to the resource speaker by condition. Each point represents the mean of each condition for each listener.*

5. Conclusions

From a technical point of view, we showed that AM was capable of a speaker identity performance at least equivalent to VC, but that both techniques still have a long way to go in achieving an output that sounds identical to the target speaker. We have only looked at one implementation of VC and there are many ways in which it could be improved, but the current study shows that it is worth investigating the use of AM in SLC in a text-dependent context. In recent years, there has been much work on HMM-based speaker adaptation which can be applied across languages, and which has become more feasible and promising with the improvement of the quality of HMM synthesis systems. It remains to be seen how AM compares to this.

The study also highlighted some of the difficulties in evaluating speaker similarity in a SLC context, partly due to listener variability and partly due to the overlap between speaker characteristics and accent. Differences between the various techniques in the way accent/language information is retained and the artefactual differences between them make it difficult for the listeners to judge speaker similarity independently from these aspects. Given anecdotal evidence that bilingual speakers sound different in their two languages, this study also raises the question of what it means to sound like the same speaker in a different language.

6. References

[1] K. Yanagisawa and M. Huckvale, “A phonetic assessment of cross-language voice conversion,” in *Proc. INTERSPEECH-2008*, Brisbane, Australia, Sep. 2008, pp. 593–596.

[2] M. Mashimo, T. Toda, K. Shikano, and N. Campbell, “Evaluation of cross-language voice conversion based on GMM and STRAIGHT,” in *Proc. EUROSPEECH-2001*, Aalborg, Denmark, Sep. 2001, pp. 361–364.

[3] D. Sündermann, H. Höge, A. Bonafonte, H. Ney, and J. Hirschberg, “Text-independent cross-language voice conversion,” in *Proc. INTERSPEECH-2006*, Pittsburgh, Pennsylvania, U.S.A., Sep. 2006, pp. 2262–2265.

[4] M. Huckvale and K. Yanagisawa, “Spoken language conversion with accent morphing,” in *Proc. 6th ISCA Workshop on Speech Synthesis (SSW6)*, Bonn, Germany, Aug. 2007, pp. 64–70.

[5] M. Huckvale, “Speech Filing System Tools (Version 4.7),” Available from http://www.phon.ucl.ac.uk/resource/sfs/, 2008, with some modifications and additions.

[6] K. Iso, T. Watanabe, and H. Kuwabara, “Design of a Japanese sentence list for a speech database,” in *Proc. Acoustical Society of Japan Spring Meeting*, Mar., pp. 89–90.

[7] J. Latorre, K. Iwano, and S. Furui, “New approach to the polyglot speech generation by means of an HMM-based speaker adaptable synthesizer,” *Speech Communication*, vol. 48, no. 10, pp. 1227–1242, 2006.

[8] P. Boersma and D. Weenink, “Praat: Doing phonetics by computer (Version 5.1.12),” Available from http://www.praat.org/, 2008.

[9] British English Example Pronunciation Dictionary (BEEP), Retrieved from http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html, 1996.

[10] Y. Stylianou, O. Cappé, and E. Moulines, “Statistical methods for voice quality transformation,” in *Proc. EUROSPEECH-1995*, Madrid, Spain, Sep. 1995, pp. 447–450.

[11] H. Ye and S. Young, “Perceptually weighted linear transformations for voice conversion,” in *Proc. EUROSPEECH-2003*, Geneva, Switzerland, Sep. 2003, pp. 2409–2412.

[12] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, pp. 37–46, 1960.

[13] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, pp. 159–174, 1977.

[14] P. Van Lancker, J. Kreisman, and K. Emmorey, “Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices,” *Journal of Phonetics*, vol. 13, no. 1, pp. 19–38, 1985.

[15] S. M. Sheffert, D. B. Pisoni, J. M. Fellowes, and R. E. Remez, “Learning to recognize talkers from natural, sinewave, and reversed speech samples,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 43, no. 7, pp. 1447–1469, Dec. 2002.