TALKING FOREIGN
Concatenative Speech Synthesis and the Language Barrier

Nick Campbell

ATR Spoken Language Translation Laboratory
ATR, 2-2-2 Hikaridai, Seika-cho, Souraku-gun, Kyoto, Japan 619-0288.
nick@slt.atr.co.jp

Abstract
This paper presents some solutions to the problem of synthesising multi-lingual speech using waveform-concatenation speech synthesis. The paper presents novel methods for deriving appropriate pronunciations for foreign words in a predominantly native-language text by use of multi-speaker synthesis, and methods for mapping the pronunciations of a foreign-language speaker onto the sounds available in the speech corpus of a native speaker so that the resulting synthesis produces speech which accurately represents the foreign words. The methods differ depending on the language-pair and on the direction of the mapping, because in the case of one-to-many phonemic mappings, high-level features can be used, but in the many-to-one case, a physical representation of the speech signal is required so that use can be made of the natural variability in speech production to select the most appropriate allophonic variant. All mappings are automatic, and the use of rule-based procedures which require human knowledge is minimised. In this way, the methods are extensible to any language combinations. Synthesised speech samples are included with the paper so that subjective evaluation of the results can be made.

1. Introduction
Many applications of speech synthesis now require the high quality of speech production that concatenative methods offer, but face increasingly multi-lingual input. To date, most applications of speech synthesis have been monolingual, and the problems of synthesising a text that contains words from more than one language are only now becoming apparent. We present details of the methods that we have used to overcome these problems and show that they are particularly suited to waveform concatenation synthesis.

Waveform concatenation synthesis takes small segments of speech from a large corpus and re-sequences them so that novel words and phrases can be formed, using the voice of the original speaker to create spoken utterances from input text. Because prosodic information is used as well as phonetic information in the selection of the waveform segments for concatenation, the use of signal processing can be kept to a minimum, and the original voice quality and speaking-style characteristics are preserved.

The translation of spoken language presents an extreme case of the need for multi-lingual synthesis, but with the growth of the internet, mixed-language texts are now becoming common. Business communications are also becoming increasingly more international, and the use of non-native-language proper-names, for people, products, and places, in a native-language text is inevitable. The problem this raises for speech synthesis is that the spellings may not be registered in the dictionary, and the letter-to-sound rules fail to predict their pronunciation satisfactorily.

The problem is confounded in the case of concatenative synthesis because the vowel-space of the original speaker may be quite different from that required to pronounce the foreign names appropriately or intelligibly. When waveforms for synthesis were largely produced by rule, in principle any sounds could be produced (although few of them were particularly natural-sounding) and the appropriate variants could be generated by rule. With concatenative methods, nearest-equivalents must be found.

2. Why multi-lingual synthesis?
Let us first consider why such multi-lingual synthesis is necessary. In the case of spoken-language translation (the main research theme of the new ATR SLT laboratory) we commonly encounter foreign names in sentences such as "Hello, I'd like to speak to Dr Yamamoto please". The name Yamamoto is not registered in the English-language pronouncing dictionary, so it is processed by the letter-to-sound rules. The first two syllables are fortuitously converted as /y uh/ and /m uh/, but the third also comes out as /m uh/ (from the misleading examples of the letter string "mo" being pronounced as in "mother", "monday" and "month", etc, instead of /m ow/ as in "Motown") and the final syllable "to" is incorrectly pronounced to rhyme with "two" instead of "toe". The resulting /y uh m uh
m uh t oo/ is an error which would confuse the listener and likely result in a misleading translation of the utterance. To confound the error, stress is then added to the penultimate syllable, rendering it closer to /y uh m uh m aa t oo/, and highlighting the mistake. Informal tests indicated that as many as half of the Japanese names pronounced through the English letter-to-sound rules will be unintelligible or unrecognisable.

Since translated utterances frequently include such personal names, addresses, product names, etc., this synthesis breakdown presents a serious problem for the intelligibility of the speech translation process. The solution we propose, extending previous work, is to use an intermediate stage of synthesis, employing the voice of a native speaker as an exemplar.

3. Previous work

In work reported earlier [3, 4] a solution was proposed for the problem of pronouncing English words using the voice of a Japanese speaker. This paper extends that work in the opposite direction to allow the voice of an English speaker to pronounce words of Japanese origin. The two problems are different. In the case of a Japanese voice speaking English, the native-language vowel space must be greatly expanded, requiring a search in the acoustic space to find the closest equivalent sounds so that the Japanese voice can reliably distinguish between words such as "cap" and "cup", or "light" and "right", when the Japanese language makes no phonemic distinction between the vowels /ah/ and /uh/ (mrpa symbols) or /l/ and /r/.

3.1. One-to-many mapping

When producing English speech using a Japanese voice database, the 15 (or so) English vowel sounds have to be appropriately reproduced from the 5 vowel locii that are available in Japanese. Even though the appropriate sounds or their close equivalents may actually exist in the speech database as allophonic variants, there is no way of accessing them directly, since the phone labels do not provide detailed enough information.

In CHATR [2], language information is stored in a knowledge-base and used in the prediction of prosodic and segmental characteristics (Figure 1). This information is used in conjunction with a knowledge-base of speaker-specific information identifying the phonetic and prosodic coverage of the speech data which is stored separately.

The solution described in [3] makes use of the knowledge about the way that a native speaker of the target language would pronounce the word sequence. The cepstral transform of the waveform data of an English speaker (generated by synthesis) is used as a target to specify the acoustic characteristics of the required speech. See Figure 2.

In the second stage of processing speech waveform segments are selected from the Japanese speaker's database by comparing their acoustic similarity to the model speech synthesised using the English speaker's voice. Thus, the knowledge-base representing the language knowledge is replaced by the cepstral vector given as input.

3.2. Many-to-one mapping

The case of an English voice speaking Japanese words is different; the English vowel space is much more diverse than that of Japanese, so simple mappings can be performed, but the prosodic control for Japanese requires tonal distinctions between phonetically similar word pairs in order to distinguish their meaning. Thus, "kaki" (pronounced like "kacky" through the English letter-to-sound rules) can mean oyster or candy depending on the relative height of the two vowels. Many such phonemically-similar word pairs exist in the Japanese language, and tonal distinctions are essential for their disambiguation.

Furthermore, prediction of the appropriate pronunciation presents a major challenge, since the majority of foreign words cannot be registered in the native-language dictionary and, for example using the case of Japanese "kaki", the default letter-to-sound rules produce output
sounding like English "cakey" or "khaki", which would be completely unintelligible to both native and foreign listeners.

4. Multi-lingual synthesis

The above discussion has focussed on the two languages English and Japanese in order to present the problem in a more readily understandable manner. In practice, the problem is not limited to just these two languages, and the methods we propose are language-independent.

To ensure this independence, we employ mapping vectors based not on direct phone-to-phone pairs, but on feature-vectors defining the phonetic space of each language. Even within a single language, there is considerable debate about the optimal phone labels to use to describe the phonetic space. The number of vowels and consonants is not absolute, and it is a matter of individual (or group) interpretation whether, for example, affricates are treated as single sounds, whether diphthongs and triphthongs are treated as compounds or as vowel sequences, etc., and semi-vowels are treated as vowels (/loy/ can be interpreted as /oʊ/ or /oʊ/ as /oʊ/, etc). The labels may be arbitrarily chosen, but the underlying phonetic features are invariant.

4.1. Mapping in the feature-space

These differences of interpretation disappear when the speech sounds are represented by a production matrix, specifying place and manner of production in the articulatory process. By representing sounds in the front-back, high-low, vocoid-contoid, pressed-open, voiced-unvoiced, etc., dimensions, we can map easily between equivalents across languages or across transcription systems. In the CHATR [2] speech synthesis system, we try to preserve database independence in order to synthesise speech from any given voice database. The dictionary for any language is a constant, producing one fixed pronunciation for any given word, but even for English, we have databases transcribed using the MRPA, DARPA, IPA, SAMPA, and OUP systems. To remain transcription-system independent, the dictionary output phone sequence is mapped into a vector of articulatory features and the database is searched accordingly.

In the case of two different languages, similar matches can usually be found; at least to the degree that their essential features can be characterised. For example, in Korean there are three degrees of aspiration on plosives. In Japanese, plosives are usually produced unaspirated, and in English, there are probably two degrees of aspiration. Such differences can be captured by the pressed-open distinction, and if using an English voice to speak Korean words, we would select an unvoiced plosive from a stressed context (more pressed) in order to represent the third degree of Korean aspiration. The phonetic systems of the languages usually differ more than the phonetic realisations of the individual phones in the speech; since as humans we all speak with the same vocal apparatus, this is not surprising. Japanese makes no phonemic distinction between /l/ and /r/ for example, but l-like sounds and r-like sounds can both be found in the actual speech production. The key is to label them in such a way that they can be accessed appropriately from the database when necessary.

4.2. Mapping in the prosodic space

By passing the non-native words through a synthesiser module for the target language, we obtain not only the phone sequence but also the predicted prosodic targets for the non-native words.

In order to map these onto the prosodic contour predicted for the utterance as a whole, we use z-score normalisation, which maps each measure onto a zero-mean unit-sd scale and allow simple addition of the predicted targets in order to derive an appropriate complex contour for the entire utterance.

We predict the English utterance-level phrasing and pitch contour for the sentence as a whole, treating the foreign name as an unspecified "flat" space, with no accentuation or stress marked, following the contour for the utterance as a whole, and leaving it to be filled in by adding the values predicted for synthesis by the Japanese speaker.

Inherent differences in pitch height and range between different speakers and speaking-styles can thus be ignored, and an accentuation pattern appropriate for the foreign word in the English sentence can be achieved.

4.3. Interface to CHATR

For the implementation of the above processes, we made use of a socket-based interface to the CHATR speech synthesiser, implemented as part of the Tcl/Tk GUI, for communication with the underlying C++ code functions. By running the synthesiser in server-mode, we were able to send messages, receive replies, and switch between speakers using a Perl script to perform the text analysis and splitting. As a condition of input, we required that the non-native words to be marked with a flag to identify their target language (in this case, a L prefix was attached to Japanese names but XML).

4.4. Flow of processing

The input text is split into native and foreign words on the basis of identifying tags. Japanese words are processed separately, and the resulting targets merged as described in Figure 3. In practice, word information is lost when phonetic targets are generated so each word in the input text is also sent individually, to allow targets for (presumably mispredicted) foreign words to be identified,
open socket to CHATR server
set language/speaker for target
send set-speaker command to CHATR
send whole text for synthesis
send generate-targets command
for each word in input text
if word is tagged as foreign
get language info from tag
set language/speaker for word
send set-speaker command to CHATR
send word for synthesis
send generate-targets command
send show-targets command
merge mapped targets
set language/speaker for target
send set-speaker command to CHATR
send merged targets to CHATR
send synth-from-targets command
close socket, go home.

Figure 3: Processing algorithm for ML-Synth.

removed, and replaced by the mapped foreign-speaker's targets.

4.5. Sample synthesis

In the following examples, we present waveforms representing each stage of the process. The first is the voice of the English speaker pronouncing the Japanese name as it was processed by the default run of the synthesiser using English-language settings. The second is the voice of a Japanese speaker (in this case a young child) pronouncing the name after default processing for a Japanese speaker. The third sample is that of the English voice using the phonemic and prosodic targets of the Japanese speaker, after mapping across both phonetic and prosodic dimensions. In all cases the resulting speech is significantly closer to the native-speaker's pronunciation, but because no signal processing was performed at all (the default for CHATR synthesis) the prosodic targets are not perfectly matched.

<table>
<thead>
<tr>
<th>NAME</th>
<th>Default</th>
<th>Japanese</th>
<th>Corrected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kurihara Kazuyo</td>
<td>kh'w.wav</td>
<td>khj.wav</td>
<td>khm.wav</td>
</tr>
<tr>
<td>Tomonaga Khiroo</td>
<td>th'w.wav</td>
<td>thj.wav</td>
<td>thn.wav</td>
</tr>
<tr>
<td>Shigeoka Fumihisa</td>
<td>sl'w.wav</td>
<td>slj.wav</td>
<td>sln.wav</td>
</tr>
<tr>
<td>Sakai Shinsuke</td>
<td>ss'w.wav</td>
<td>ssj.wav</td>
<td>ssn.wav</td>
</tr>
<tr>
<td>Nagisa Yasue</td>
<td>ny'w.wav</td>
<td>nyj.wav</td>
<td>nyn.wav</td>
</tr>
<tr>
<td>Mine Toshiohi</td>
<td>mtl'w.wav</td>
<td>mtlj.wav</td>
<td>mtln.wav</td>
</tr>
<tr>
<td>Tani Hiroaki</td>
<td>thl'w.wav</td>
<td>thlj.wav</td>
<td>thln.wav</td>
</tr>
<tr>
<td>Watanabe Hideyuki</td>
<td>whl'w.wav</td>
<td>whlj.wav</td>
<td>whln.wav</td>
</tr>
<tr>
<td>Hirono Mai</td>
<td>lm'w.wav</td>
<td>lmj.wav</td>
<td>lmn.wav</td>
</tr>
<tr>
<td>Yokoyama Kiyomi</td>
<td>yk'w.wav</td>
<td>ykj.wav</td>
<td>ykn.wav</td>
</tr>
</tbody>
</table>

5. Summary

Speech synthesisers need to be able to speak more than one language. With parametric synthesis the problem is simpler, but the speech quality correspondingly poorer. With concatenative speech synthesis using minimal signal processing, it would be optimal to record samples of speech from a poly-lingual speaker, but such people are rare and perhaps non-existent for arbitrary language pairs.

The proposed methods both involve copying a native-speaker's production, but in the one-to-many case, a stage of intermediate processing can be eliminated. When mapping from a language with many vowels to a language which has inherently fewer, it is necessary to produce an acoustic target exemplar to ensure an appropriate match by searching through allophonic variations of the same phonemic sound, but in the reverse direction, a feature-based search in the phonemic space appears to suffice, providing that the prosodic characteristics can be preserved.

6. Conclusion

This paper has presented a solution to the problem of synthesis across languages, by presenting mapping techniques which enable the voice of one speaker to be used for the generation of speech containing words from multiple languages. It used as an example the pronunciation of Japanese names by an English speaker, extending earlier work that enabled the pronunciation of English words by a Japanese speaker.

A program has been implemented to demonstrate the performance using an English voice speaking Japanese names. Sample speech synthesised by this process confirms that pronunciation improves considerably under the proposed method.

The pronunciation of the target sequence is limited by the sounds available in the database for the speaker of the native language, giving the perception of a "foreign-accent" in the synthesised speech, but we believe that this is a natural style of pronunciation. The fact that the native speaker is perceived to be fluent in the foreign language is an illusion that lends merit to this method of speech synthesis. The phonetic sequence and the prosodic patterns are close to those of a foreign speaker, but the vowel colouring is that of a native speaker, thus rendering the speech acceptable to native and foreigner alike.

7. References