PRONOUNCING FOREIGN TEXT TECHNIQUES FOR CROSS-LANGUAGE SPEECH SYNTHESIS

Nick Campbell,
"ATR Spoken Language Translation Labs

ABSTRACT

This paper describes a method for synthesising foreign words using the voice of a non-native speaker in concatenative speech synthesis. Rules are described for mapping across phone sets between languages, and for selecting speech units close to the native-language. We focus on the case of an English voice producing Japanese words as part of a predominantly English sentence. A procedure utilizing prosodic and phonetic targets predicted for a Japanese speaker is proposed.

1. INTRODUCTION

There is a growing need for speech synthesis to be multi-lingual, and for speech synthesisers to speak in more than one language, without changing the voice of the speaker. This is a challenge for concatenative synthesis, which uses the voice of a pre-recorded speaker as a source of units. When recording source-units for a speech synthesis database, it is difficult to find a reference speaker who has the same level of fluency in several languages, so a mapping must be found between the sounds of the speaker's native language and those of the target language, in order for the voice to be used multi-lingually.

As with human performance, a slight "foreign accent" may be acceptable when pronouncing any non-native words, but the grammar and pronunciation should be as close to that of the target language as possible. The stress, accent patterns, timing, and rhythm of the foreign words should be preserved in order for their sense to be intelligible to listeners of both languages.

2. MULTI-LINGUAL SPEECH

There are applications for cross-language speech synthesis technology in educational and entertainment areas, but it is in spoken-language translation where the need is most urgently felt. When translating dialogue speech, in particular, it is common to find use of words, especially proper names (forenames surnames, city names, hotel names, addresses, etc.,) from more than one language in the same sentence.

Recent advancements in speech recognition have solved many of the problems of proper-name recognition [1], but open up related problems as these "foreign" names are required to be pronounced by the speech synthesis module and are passed in plain text as part of a sequence of native-language words.

「外国語混合文対応型多言語音声合成方式の開発」

・ ニック キャンベル（ATR音声言語通信研究所）

2.1. Spoken Language Translation

In translated dialogues, especially in the travel domain, it is common for foreign names to be included amongst the target-language words in the translated text, as the following example illustrates:

e.g. 1, "My name is Yamamoto and I'd like to reserve a room at the Washington Teikoku Hotel for three nights."

In order to determine the pronunciation of English words in CHATR [2], a dictionary is used in conjunction with letter-to-sound rules. Only when the pronunciation of a word cannot be predicted by rule, is its pronunciation stored in the dictionary.

The problem in synthesizing foreign-language proper-names arises from the fact that they are too numerous and unpredictable to pre-register in the native letter-to-sound dictionary and that they are not usually well handled by the native-language letter-to-sound conversion rules, causing serious problems of intelligibility in the synthesized speech.

In the case of the name "Yamamoto" in the example above, the English pronunciation rules produce /y a m a m aa t uu/ instead of /y a m a m o t o/ because the letter sequence [mo] is parsed (mistakenly) as equivalent to that of "mother" or "month", producing /m aa/, and the [to] sequence is parsed as for the English word "to", being pronounced as /t uu/. As is often the case, the consonant components are well matched, but the vowel variant is wrong in 2 out of 4 cases.

2.2. Pronouncing non-native words

Previous work [3] described how a speech signal can be used instead of parametric targets as a unit-selection criterion for multi-lingual concatenative synthesis, so that the appropriate vowel variants from the database of a Japanese speaker can be selected to be closer to those of an exemplar English speaker.

This use of an intermediate stage of synthesis, using the voice of a native-speaker to produce the spectral target, allows the distinction of e.g., [cup] and [cap] even when the choice is only from the "ア行" section of the Japanese speech database. However, in the case of pronouncing Japanese words using an English voice, the richer vowel set of the English language allows a one-to-many choice to be made, and the simpler task of checking the phone-sequence predictions will suffice, as long as the target-language prosody is respected.

We therefore propose a variant of two-stage synthesis in order to pronounce Japanese words using the voice of an English speaker.

3. PROSODY-PHONE MAPPINGS

If the input text is tagged by the translation component to distinguish between the translated and non-translated words in the utterance to be spoken, e.g., using XML markup, then we can first use a Japanese processing module to generate a set of phone and prosodic targets, and then, on the basis of these, correct the English pronunciation and transfer across the prosody predicted for the non-native words.

e.g. 2, My name is <foreign language=Japanese> Yamamoto </foreign> <native language=English> and I'd like to .....

By comparison with the phone sequence predicted for Japanese, we can substitute the closest English vowel (in this case /oh/ for the mis-predicted vowels /aa/ and /uu/), and then map the pitch power, and duration targets directly in a one-to-one manner from the Japanese, as targets for the English phones.

3.1. Phone-based distance measures

The phone mappings in CHATR are performed by way of phonetic feature vectors, when not explicitly declared. The speech sounds are defined by a matrix of place and manner parameters, so by the use of simple Boolean distance metrics an equivalent sound pair can be found across any language combination. Tables 1 and 2 illustrate some of the features used (here showing phone-label, sonorance, length, height, fronting, rounding, manner, place, voicing, and internal label).

Table 1: phone features for Japanese vowels (not complete)

a	+	s	3	2	-	0	0	+	vow_a
I	+	s	1	1	-	0	0	-	vow_I
u	+	s	1	3	-	0	0	+	vow_u
e	+	s	2	1	-	0	0	+	vow_e
o	+	s	2	3	-	0	0	+	vow_o

Table 2: phone features for English vowels (not complete)

ax	+	a	2	2	-	0	0	-	schwa
axr	+	s	2	2	-	1	r	-	rho-schwa
aw	+	d	3	3	+	0	0	+	dipth-aw
ow	+	1	2	3	+	0	0	+	vow_ow
oh	+	s	2	3	-	0	0	+	vow_oh

3.2. Prosody mapping

The prosody predicted by the English accentuation modules is not likely to be appropriate for the pronunciation of Romanised Japanese letter sequences. In this case, we can take as example the prosodic parameters predicted for the Japanese speaker and map them onto the phonetic sequence to be generated by the English voice. The use of normalised values for duration, power, and pitch (i.e., their z-score converted values) allows direct mapping across speakers on a phone-by-phone basis

Table 3: Predicted phones, durations, and pitch for Japanese and English versions of the same Japanese word:. The final phone mappings are shown in the center.

Y	73	134	=	y	87	144
a	36	168	=	ae	68	151
m	32	163	=	m	62	140
a	62	183	=	ae	37	137
m	32	164	=	m	82	122
o	118	137	o h	aa	134	126
t	78	126		t	88	108
o	128	116	= o h	uu	121	93

This algorithm differs from that reported in [3] by not requiring an acoustic target for the Japanese-to-English direction. The text of the Japanese word sequence is first synthesized by rules for a Japanese speaker (but without use of the unit-selection and waveform concatenation stages) and the prosodic targets are then mapped directly onto the English phones. For this mapping stage, it is necessary to first check that the English letter-to-sound rules have produced an appropriate sequence of phones. This requires dynamic string-matching between the phone sequence predicted by the Japanese kanji-to-romanji modules and those predicted by the letter-to-sound module, using the feature-based distance measures for the vowel mappings.

4. SUMMARY

An algorithm has been presented which performs the mapping of pronunciation and prosody between two languages. The difference in the size of the vowel-sets of English and Japanese requires different processes depending on the direction of the conversion. The algorithm enables high-quality synthesis of words from more than one language to be performed using the voice of a speaker of (in principle) any language. It has been tested for the pronunciation of Japanese proper names by English speakers, but is also applicable to longer sequences of non-native text and renders language independent of speaker.

5. REFERENCES

1. "A Hierarchical language model incorporating class-dependent word models for OOV words recognition" Tanigaki, K., Yamamoto, H., Sagisaka, Y. pp 123-126 in Proce ICASSP 2000.
2. Campbell, W. N. and Black, A. W. CHATR a multi-lingual speech re-sequencing synthesis system. *Technical Report of IEICE* SP96-7, 45-52, 1996.
3. Campbell,W.N., "Foreign-Language Speech Synthesis", Proceedings ESCA/COCOSDA 3rd Speech Synthesis Workshop, Jenolan Caves, Australia 1998/11/26.