FOREIGN-LANGUAGE SPEECH SYNTHESIS

Nick Campbell
ATR Interpreting Telecommunications Research Labs.
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, JAPAN
nick@itl.atr.co.jp, www.itl.atr.co.jp/chatr

ABSTRACT

This paper describes a method of concatenative speech synthesis for producing speech in a language other than that of the database speaker. In certain applications, such as interpreted dialogues or multi-lingual e-mail, it is necessary to synthesise words that are foreign with respect to the language of the main text. In this case, rather than switch voices, we show that the use of an intermediate stage of synthesis improves the pronunciation and prosody of the output speech.

1 INTRODUCTION

Sadaoki Furui once made the challenging observation that computer processing of speech should not be limited to simply reproducing human abilities, and that computers should be expected to offer something above and beyond the levels of human performance [2].

In pursuing multi-lingual speech for the CHATR concatenative synthesis system[1, 5], we are taking up Furui's challenge and offering human speakers the ability to extend their performance by appearing to speak in foreign languages that they may not in fact know.

Our main purpose for this work is to produce a component for an interpreting telecommunications system[6] in which the input speech is recognised, translated, and synthesised in the target language, using the voice of the original speaker. However, this component has broader applications in view of the need to pronounce the foreign words that appear in a multi-lingual text, such as when synthesising e-mail or HTML pages from the world-wide-web.

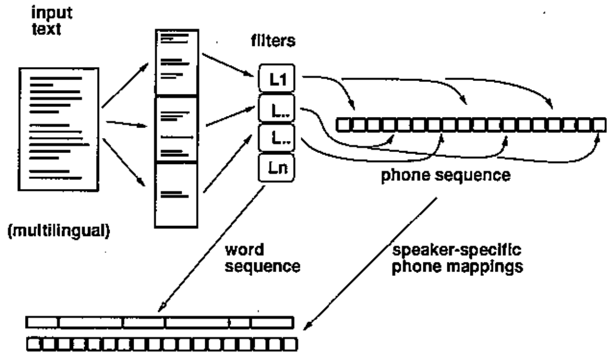


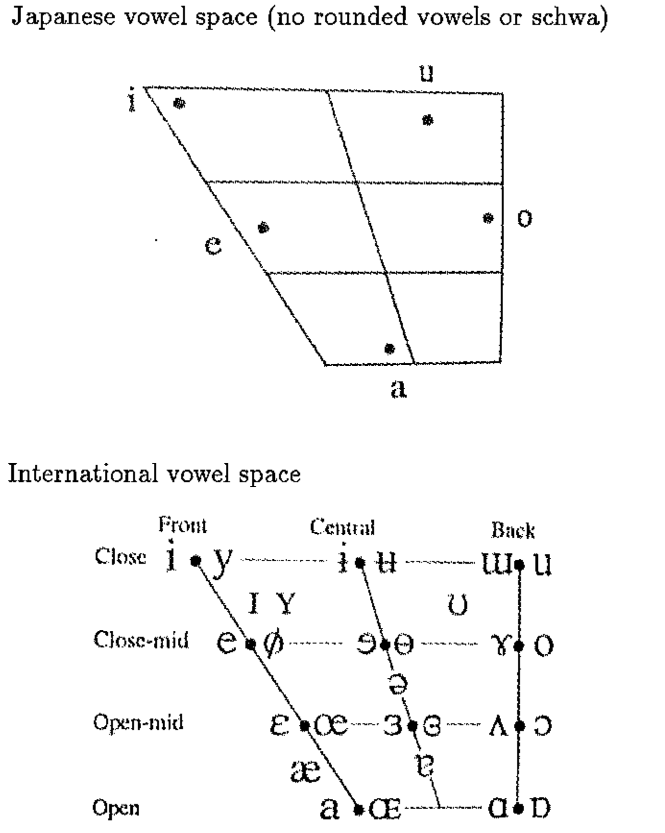
Figure 1: Mapping the phone sequences for each language represented in the text onto the phoneset used by the speaker for the synthesis.

Table 1: Mapping table from English to Japanese

ENG	JPN	ENG	JPN	ENG	JPN
ax	u	axr	a	aa	aa
ao	a	ah	a	ay	ai
aw	au	ae	a	ea	ea
ia	ia	ua	ua	el	l
en	N	er	a	eh	e
ey	ei	iy	ii	ih	i
uh	u	uw	uu	em	m
oh	o	ow	oo	oy	oi
y	y	r	r	l	r
m	m	n	n	ng	N
nx	N	jh	j	ch	ch
zh	z	sh	sh	th	s
dh	z	p	p	b	b
d	d	dx	d	t	t
k	k	g	g	f	f
v	b	z	z	s	s
hh	h	w	w	sil	#
brth	@	laugh	lx	outbr	X

Monolingual speech synthesis using recognisable human voices has already been demonstrated, and we have shown that with a sufficient source corpus CHATR is capable of reproducing the voice and speaking style of a given speaker with high fidelity[3]. This is achieved by the concatenation of phone-sized waveform segments without recourse to signal processing for the modification of prosody. Voice quality is preserved by selection and concatenation of units that are naturally close to the desired prosody and which therefore do not need (potentially damaging) modification. [SOUND 0024.01.WAV][SOUND 0024.02.WAV][4]

In this paper, we describe a method for a) selecting a sequence of segments that best match the sounds of the target speech through the use of a mapping vector, and b) using as an intermediate synthesis stage the voice of a native speaker of the target language to provide an objective physical model with which to constrain the unit selection. Samples of the speech produced using this method can be heard on our web pages[1].


Japanese vowel space (no rounded vowels or schwa)

International vowel space

Figure 2: Reproduced (with thanks) from the Journal of the International Phonetic Association.

2 MULTI-LINGUAL TEXT

With the increasingly international nature of the working environment, there is frequent exposure to multilingual texts or texts that are written predominently in one language but that contain words from another. We can distinguish two classes of such texts; those that are 'parallel' bilingual, or multi-lingual, and which present translated versions of the same message in each language, and those which are 'embedded' multi-lingual, which have only one message that contains words of more than one language. ("In Japanese 「音声合成」 means 'speech synthesis'." is an example of the latter.)

It is a matter of policy how to synthesise parallel multi-lingual texts, and in many cases it is probably sufficient to produce speech only in the preferred language of the listener, ignoring the parallel text completely. However, in embedded multi-lingual texts, the foreign words are part of the message and must be synthesised for every listener. In such a case, the synthesiser needs to be capable of processing text and generating speech in more than one language. Earlier versions of CHATR used two voices for this task, one for each language, but we found switching voices in mid-sentence to be unaccpatable.

Figure 1 illustrates the text-level processing flow for multilingual synthesis. In many language combinations, the language type can be determined from the encoding of the computer symbols that represent the text. In our case, Japanese (EUC/JIS), English (ASCII), and Korean(EUC/KIS) are common and can be easily distinguished. Once the different languages have been recognised, then the grapheme-to-phoneme filters for each language can be applied to produce a phonetic rendering of the utterance. The output from the different language filters is then recombined to form a complete sequence, which can then be mapped onto the phonetic space of the output speaker.

For a parametric synthesiser, rules would be required for any additional sounds that are not usually found in the language of the speaker, but in the case of concatenative waveform synthesis, a mapping must be performed so that the native-language sounds which make up the speaker's database can be re-used to form the closest approximation to the desired target sequence in the foreign language.

For Japanese, for example, it will be noticed from Figure 2 that, in contrast to English, the central vowel space (schwa) is unused. There is no phonemic reduction in this language but there is natural variety in the speech which, if correctly identified, can be used as a substitute. This paper describes ways to find the best segments.

When producing English speech using a Japanese voice database, the 15 (or so) English vowel sounds have to be somehow mapped onto the 5 vowel locii that are available in Japanese. Table 1 presents an example of such a mapping vector from (machine-readable) English into Japanese. We can see that pairs of words like 'cap' and 'cup', and 'lice' and 'rice' become impossible to distinguish unless further clues are available from the text.

3 MULTI-LINGUAL SPEECH

When a Japanese person speaks in English, unless he or she is particularly fluent, the range of variability in the resulting vowel space will probably be closer to that of the mother-tongue than to that of a native-speaker of English. This is one of the causes of 'foreign accent'. The restricted range of prosodic variation, in accordance with mother-tongue patterns, is another. The phonetic and prosodic habits of our first language can be very difficult to overcome when speaking in a second language.

However, many people can successfully communicate in foreign languages without really departing far from the prosodic and phonemic spaces of their native lan-

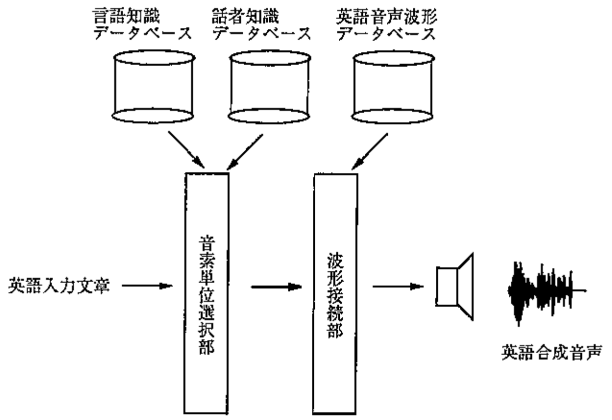


Figure 3: Two knowledge bases and a speech database provide the processing power in CHATR

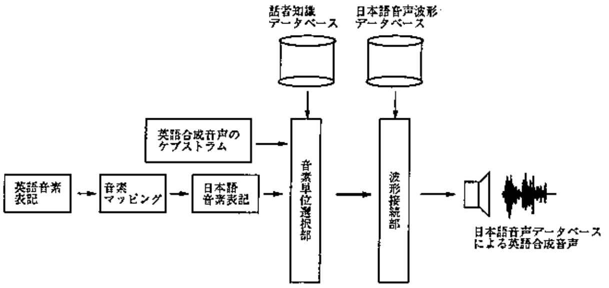


Figure 4: The language knowledge base is replaced by a cepstral target given as input.

guages. They do this by re-sequencing their own familiar speech sounds in an order that is appropriate to (at least) the lexis of the target language.

3.1 Phone Mapping

In CHATR synthesis, each source-speech database is labelled using a phone-set determined for the speaker and dialect of the language in which the recordings were made. By mapping from the phone sequence predicted by the individual grapheme-to-phoneme components to the phone-set used to label the source database speech, we can produce foreign-language speech using the voice of any speaker.

In the following examples we use the voice of a small Japanese child to speak in English ([SOUND 0024.03.WAV][SOUND 0024.04.WAV] greetings) and Korean ([SOUND 0024.05.WAV] [SOUND 0024.06.WAV] explaining details of the technical processing within CHATR). Although the phone sequence is appropriate, the fine phonetic detail is more suitable for Japanese, and the prosody is closer to the patterns of Japanese than Korean or English. The effect is that of a small Japanese child speaking fluently in the foreign languages. The speech is intelligible, but accented. It is probably a good representation of the way a non-native speaker would produce the utterance if given some coaching in the foreign language, but is not adequate for synthesis applications if the listener has had no experience of listening to such accented speech.

3.2 Two-stage Language Mapping

One reason the phone mapping technique does not produce ideal speech in the foreign language is that, even though the appropriate sounds or their close equivalents may exist in the speech database, it has no way of targeting the phonetic detail of the speech sequence. In order to find a sequence of segments that more closely represent the way that a bilingual speaker or a native speaker of the language would produce the utterance, we need to select more finely within the natural variation of the speaker's data. For this, we need information that is more detailed than the phone labels can provide.

Figure 3 shows the typical sequence of processing within CHATR. Language information is stored in a knowledge-base and used in the prediction of prosodic and segmental characteristics. 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 being described in this paper is illustrated in Figure 4, which shows a secondary stage of synthesis that makes use of the knowledge about the way that a native speaker of the target language would pronounce the word sequence (synthesised as shown in Figure 3).

The waveform data of the first speech sequence (or its cepstral transform) is taken as a model to specify the acoustic characteristics of the desired speech. In the second stage of processing (Figure 4) we select speeech waveform segments from the non-native speaker's database by comparing their acoustic similarity to the model speech synthesised using the native speaker's voice. Thus, in Figure 4 the knowledge-base representing the language knowledge is replaced by the cepstral vector given as input.

For each phone of the synthesised target speech, we then perform a cepstral-based scoring of every candidate phone in the Japanese speaker's database that has the appropriate phone label after mapping. The cepstral distances are computed as follows: For each frame of the target cepstrum vector

$$c_{1,t}(0), c_{1,t}(1), c_{1,t}(2), \ldots, c_{1,t}(M_t)$$

and each frame of the candidate unit cepstral vector

$$c_{2,t}(0), c_{2,t}(1), c_{2,t}(2), \ldots, c_{2,t}(M_u)$$

the distance is calculated as the square of the individual differences

$$d(t) = \sum_{k=0}^{M_t} (c_{1,t}(k) - c_{2,t}(j))^2$$

and an overall score determined for each candidate.

$$d = \frac{1}{M_t} \sum_{t=0}^{M_t} d(t)$$

The sequence of candidates having the lowest cepstral-distance scores is passed as input to the waveform concatenation component.

4 EVALUATION

A full evaluation of this algorithm has not yet been performed as we are currently refining the mapping table by allowing for multiple unit candidates per input vowel. However, an informal listening test using MOS scores (best=5, worst=1) yielded the results shown in Table 2.

Table 2: MOS scores for each method of synthesis

target	phonemic	phone-map	cep target
interesting	intrustiN	2	3
forever	furebu	1	1
trouble	trabr	2	2
difficult	difikurt	2	2
rump	ramp	3	4
lamp	ramp	3	4
expect	ikspekt	2	2
shopping	shopiN	3	4
deep hole	diip hoor	3	4
your family	yo famuri	2	3
Average		2.3	2.9

Overall quality was not good when using the mapping table alone, but improved considerably when use was made of the intermediate native-speaker cepstral vectors to refine the selection. The test word 'forever' presented the most difficulty as it was pronounced with two syllables, both fully reduced. The words 'lamp' and 'rump', phonemically identical when transliterated into Japanese, improved considerably, enabling correct distinction of each. Improvements were noted in both the phonetic realisation and the prosodic contours of the synthesised speech when using cepstral-based selection.

5 CONCLUSION

To reduce the 'accentedness' of the mapping-based synthesis, we adopted a two-stage process, first synthesising the target speech using the voice of a native speaker of the target language and then using the acoustic waveform (or its cepstral representation) as a physical target for the selection of speech segments from the prestored voice database of the input speaker by minimising a physical distance measure.

This use of a physical target for unit selection is not feasible in monolingual synthesis since, by definition, if the utterance existed somewhere in a suitable form there would be no need to synthesise it. However, by making use of an intermediate target we can narrow down the selection of speech segments to match the spectral characteristics of the native, thereby making use of the natural variation in production that could not be accessed through label information alone.

The effectiveness of the method relies on correctly identifying the appropriate variant from amongst the different articulations of each speech sound in the source database. The method as it is currently implemented is limited in that the candidate segments are first prefiltered by a many-to-one mapping table. Future work includes training a many-to-many mapping table based on the acoustic distances between units in a bilingual speaker's database.

Many speech synthesisers are capable of multi-lingual output but for their prosodic manipulation most make use of signal processing that results in a mechanical-sounding voice which is often no longer recognisable as that of the original speaker. Since CHATR produces speech using the recognisable voices of known people, it offers the potential to extend a person's apparent abilities into the realm of multi-linguality. By offering this ability to the voices of young children, we claim to have achieved Furui's goal.

Acknowledgement

I would like to express my thanks to Ken Fujisawa and Norikazu Tamura for their assistance, to Masanobu Abe for suggesting the use of a bilingual speaker, and to ATR-ITL for the support that enables this research to be performed.

References

[1] http://www.itl.atr.co.jp/chatr.
[2] Sadaoki Furui, closing panel session at ICSLP-96, Philadelphia.
[3] Scientific American Frontiers (PBS, April '98) Digital Alan.
[4] http://www.itl.atr.co.jp/chatr/j_tour/yuto2.html.
[5] W. N. Campbell, "CHATR: A High-Definition Speech Re-Sequencing System", Proc 3rd ASA/ASJ Joint Meeting, 1223-1228, Hawaii, 1996(12).
[6] Toshiyuki Takezawa, Tsuyoshi Morimoto, Yoshinori Sagisaka, Nick Campbell, Hitoshi Iida,Fumiaki Sugaya, Akio Yokoo, Seiichi Yamamoto: "A Japanese-to-English Speech Translation System: ATR-MATRIX," Proc. ICSLP98, (December 1998).