Investigating Prosodic Modifications for Polyglot Text-to-Speech Synthesis

Péter Olaszi, Tina Burrows, Katherine Knill {peter.olaszi, tina.burrows, kate.knill}@crl.toshiba.co.uk
Speech Technology Group, Cambridge Research Laboratory, Toshiba Research Europe Limited, 1 Guildhall St, Cambridge, CB2 3NH, United Kingdom

1. Aim

Test our ***assumptions*** that for synthesising mixed-lingual texts using a polyglot text-to-speech synthesis (TTS) system with a primary language:

1. native prosody in foreign inclusions will produce more natural and intelligible speech;
2. as the length of inclusion increases, the importance of applying native prosody increases.

2. Introduction

There is a demand for TTS systems that can handle mixed-lingual texts. Examples: automated cinema booking systems reading foreign film titles; in-car navigation systems that pronounce foreign place names.

***Multilingual solution**:* Each text portion of a different language is synthesised by a corresponding monolingual TTS system.

- The voice is different for each language
- Switching voices is very difficult to listen to

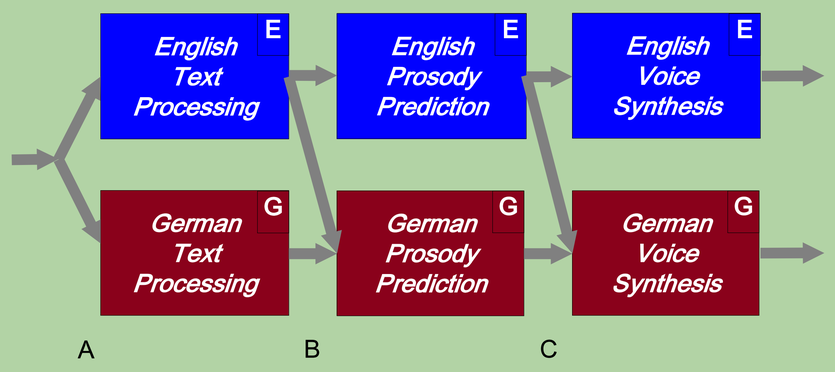
***Polyglot approach**:* The same voice and speaker identity is maintained throughout.

- Single voice is much more natural
- Problem: maintain voice quality over all languages

In our experiments we used an existing monolingual speech database, and mapped foreign phones to the existing ones without modifying the speech data.

***Question:** If we have a sentence in a "base" language with foreign inclusions, how much "foreign" processing is required in the foreign parts?*

3. System configuration: a hybrid English–German TTS system



4. Language crossover points

Only the German voice was considered for the mixed-lingual voice, so no crossovers from German to English shown.

***A**: Language selection*: Separation of the input text portions according to their language.

***B**: Crossover after text processing*: The phone sets, part-of-speech and syntactic role tags are language specific. They are mapped to the input of the German prosody prediction module.

***C**: Crossover after prosody prediction*: Only the phones need to be mapped from English to German.

5. Mapping of phones and grammatical information

English phones missing from the German phone set were substituted by phonetically similar German phones. The mapping was done by manually created mapping tables.

One-to-one mapping was used for the consonants and monophthongs.

One-to-many phone mappings were applied to convert English diphthongs into a pair of German monophthongs (e.g. /ei/ → /e/ + /i/).

Part-of-speech tags and syntactic roles were also mapped by table lookup.

6. Method of evaluation

- Stimuli: 4 text input categories and 4 system configurations were combined in 11 setups – resulting in 110 stimuli altogether.
- The stimuli set was split into two blocks of 55 utterances each, to keep one listening session below half an hour.
- Subjects: 13 native speakers of English and 13 native speakers of German, all having good or excellent knowledge of the other language.
- Questions about intelligibility, naturalness of prosody and acceptability. Mean opinion score on a scale of 5 (1 worst, 5 best).

7. Setups for creating stimuli

System Configuration	Text Analysis	Prosody Prediction	Voice	Inputs categories	Number of stimuli
EEE	English	English	English	*Eng*	10
EEG	English	English	German	*Eng*	10
EGG	English	German	German	*Eng*	10
GGG	German	German	German	*Eng, Ger, Inc, Mix*	4 × 10
EEG/GGG	English/ German	English/ German	German	*Inc, Mix*	2 × 10
EGG/GGG	English/ German	German	German	*Inc, Mix*	2 × 10

EEG/GGG and EGG/GGG are hybrid configurations for mixed-lingual input, where text processing and prosody prediction is used depending on the language of the text fragment.

8. Input text categories

Name	Text categories, 10 sentences in each	Average sentence length in words		
		English	German	Total
Eng	Purely English sentence	12.4	–	12.4
Ger	Purely German sentence	–	12.3	12.3
Inc	German sentence with short English inclusions	3.2	7.8	11.0
Mix	Mixed-lingual sentence with English and German clauses	8.8	6.6	15.4

9. Examples of input sentences

Eng	*I would be grateful for an indication of your rates.*
Ger	*Es tut uns leid zu hören, daß die Ware beschädigt ist.*
Inc	*Am Trafalgar Square, nahe der Lord Nelson Statue, wird jedes Jahr ein riesiger Weihnachtsbaum errichtet.*
Mix	*Gravitation is not responsible for people falling in love – hat einmal Albert Einstein gesagt.*

10. Results

Setup		Intelligibility		Naturalness of Prosody		Acceptability	
Text	*System*	*English*	*German*	*English*	*German*	*English*	*German*
Eng	EEE	4.21	4.35	2.98	4.16	3.06	4.13
	EEG	3.06	3.15	2.92	3.24	2.58	2.87
	EGG	2.69	2.66	2.42	2.58	2.17	2.25
	GGG	1.85	1.92	2.17	2.39	1.51	1.48
Ger	GGG	4.38	4.59	3.73	3.62	4.07	4.28
Inc	EEG/GGG	3.82	3.58	3.48	3.36	3.56	3.44
	EGG/GGG	3.88	3.65	3.53	3.32	3.62	3.44
	GGG	3.43	3.34	3.16	3.12	3.19	3.12
Mix	EEG/GGG	4.01	3.92	3.28	3.57	3.55	3.65
	EGG/GGG	3.73	3.48	3.03	3.02	3.23	3.23
	GGG	2.75	2.47	2.79	2.83	2.32	2.12

Green lines: Statistically significant mean differences in adjacent cells
Red lines: Statistically not significant differences in scores.

11. Discussion

1. Clear preference of both German and English native speakers for the use of language specific processing
   - For all systems with English text input included there was a big improvement in the evaluation scores for intelligibility and acceptability from the German only system (GGG) to the systems with English processing.

▸ ***First assumption verified***

2. For mixed (*Mix*) and purely English texts (*Eng*) using English prosody in polyglot synthesis improved the scores
   - For the *Eng* and *Mix* inputs all but one mean differences in adjacent rows of EEG and EGG were statistically significant when tested by paired t-test with a 95% confidence interval.
3. For short inclusions of a few words (*Inc*) there was no clear preference between the English and German prosody.
   - None of the mean differences in adjacent rows of EEG and EGG were statistically significant.
   - For short foreign inclusions, either native or foreign prosody may be used, depending on the convenience of the target application.

▸ ***Second assumption verified***

4. German native speakers ranked the prosody and acceptability of the English monolingual system (EEE, *Eng*) higher than the English native speakers.
   - For EEG configurations on *Eng* texts, German listeners again tended to give higher prosody and acceptability scores than the English subjects.
   - This suggests that the German listeners were more forgiving of errors heard in the other language.

12. Conclusions

In polyglot speech synthesis

- Using native prosody as well as native text processing in foreign inclusions produces more natural and intelligible speech
- The importance of applying native prosody increases as the length of text inclusion increases.
- Adding foreign prosody will not degrade short inclusions of foreign text