Investigating Prosodic Modifications for Polyglot Text-to-Speech Synthesis

Péter Olaszi, Tina Burrows, Kate Knill

Speech Technology Group, Cambridge Research Laboratory,
Toshiba Research Europe Limited, 1 Guildhall St, Cambridge, CB2 3NH, UK
{peter.olaszi, tina.burrows, kate.knill}@crl.toshiba.co.uk

Abstract

This paper investigates the need for applying English prosody when synthesising English portions of mixed English/German texts using a German-based polyglot text-to-speech (TTS) synthesis system. The polyglot system is based on a monolingual German TTS system, which uses a phone mapping from English to German to synthesise English texts. Two systems with varying degrees of assimilation to English are compared, one in which prosody is derived from the German monolingual system, and one in which the prosody is derived from an equivalent English monolingual system. The naturalness of the different prosody approaches and overall intelligibility and acceptability of the polyglot systems is assessed by native bi-lingual speakers of both English and German, on German texts containing varying lengths of English inclusions, and on complete texts in English. The results show that both German and English subjects preferred English prosody for longer English inclusions or complete English texts, but had no preference for short inclusions.

1. Introduction

In everyday speech applications, there is an increasing demand for text-to-speech synthesis systems that handle mixed-lingual texts, in which a text in a primary language includes words, phrases or paragraphs in other 'foreign' languages. Example applications include automated cinema booking systems, where foreign film titles may feature, and in-car navigation systems which have to pronounce foreign place names. In a 'multilingual' solution to this problem [1], each text portion in a different language is synthesised by a corresponding monolingual TTS system. Unless each of the systems is trained on the voice of the same multilingual speaker, the different language systems will have different voices. While this may be acceptable for complete texts in only one language at a time, it is less appropriate for mixed-lingual texts. To handle such texts, a 'polyglot' solution [1] may be more appropriate, where the same voice and speaker identity is maintained throughout. All multilingual and polyglot systems are faced with the challenge of identifying the required language. In cases where the system is based on a specific primary language, there is an additional challenge of determining the degree of assimilation of a 'foreign' language which is required.

As previously mentioned, one solution to a single voice for multiple languages is to train multilingual or polyglot systems on a multilingual speaker [1]. However, recording a multilingual corpus is expensive and not scalable, since it limits the scope of the system to the languages covered by the chosen speaker. Another approach is to synthesise using an unmodified monolingual TTS system with a voice native to the primary language of the mixed-lingual texts. Although simple, such a system suffers from gross errors in pronunciation and text normalisation of the foreign language and rapidly becomes unintelligible for longer, more complex foreign inclusions or entire texts in a foreign language [2]. Most previous work on polyglot solutions has focused on assimilation at the linguistic level, by applying a 'foreign' linguistic model to a monolingual TTS system with a voice native to a chosen primary language. The linguistic model may include text analysis and normalisation, a grapheme-to-phoneme model and a mapping between the phone set of the foreign language and the primary language of the TTS system [2][3][4]. More complex approaches have utilised cross-language voice conversion techniques [5] and adaptation of a 'generic' (language and speaker independent) polyglot voice trained on a combination of data from multiple languages and speakers to a specific native voice for synthesising non-native languages [6].

Assimilation at the linguistic level is fairly successful for phonetically similar languages [4]. The resulting foreign synthesized speech is more intelligible than that from an unmodified non-native monolingual system but still retains a degree of accent of the primary language to a native speaker of the second language. This is partly due to non-exact correspondence between phone sets but also to differences in prosody. Some degree of foreign accent may be acceptable to a native speaker of the primary language, for example for strongly assimilated loan words or short inclusions, and has been shown, in some cases, to actually improve acceptability [7]. However, the 'foreign' accent may hinder intelligibility for longer inclusions or entire texts in a second language. In such cases, full assimilation to the second language may be more appropriate and improve the quality of the synthesis [3].

In this paper, the need for assimilation of prosody in the synthesis of mixed-lingual texts is investigated for an English-German polyglot system using German as the primary language. Section 2 describes the architecture of the polyglot system and section 3 the experimental setup. Experimental results and their analysis are presented in sections 4 and 5.

2. System setup

Figure 1 shows the architecture of a German-English polyglot text-to-speech (TTS) system as a sequence of three basic language-specific processing modules: text processing,

prosody prediction and speech signal generation. As the focus of these experiments was on how a German voice can produce German, English and mixed-lingual utterances, German was taken as the primary or 'native' language for synthesis of mixed-lingual texts. Only crossovers between modules from English to German are considered here, however, the architecture could simply be revised to include crossovers in the other or both directions.

The text processing module contains sub-modules for pronunciation prediction (including word stress assignment) and syntactic analysis. Each sub-module is monolingual. The phone sets, part-of-speech (POS) and syntactic role tags output by the text processing module are language specific. They therefore need to be converted if used in subsequent modules in an alternative language. The prosody prediction module contains sub-modules that predict prosodic effects such as prosodic chunks, pausing, accenting, duration and pitch. The output of prosody prediction is a phone sequence with durations for each phone and pitch values for each frame. The latter two items are language independent so only the phones need to be mapped at a crossover following prosody prediction. Finally the voice synthesis module converts its input into synthesised speech.

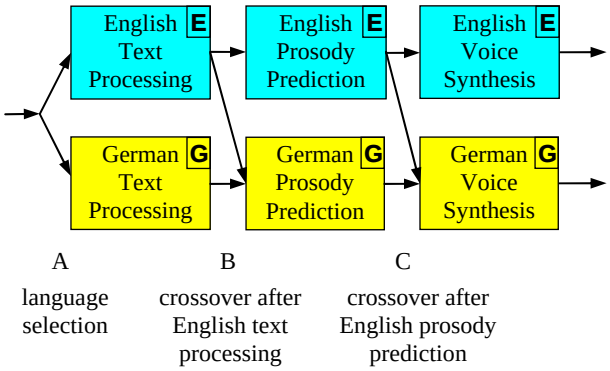

Figure 1: Polyglot TTS architecture showing crossover points from English to German processing.

For processing of English language fragments, two crossover points from English to German processing were considered, as shown in Figure 1. At both crossover points, as the German speech database contains only German phones, and none of the English-specific ones, the English speech sounds have to be mapped to the German speech sounds. This was approximated using a mapping table. The substituting phones were selected by the phonetic-phonological similarities of the speech sounds' features. A 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/). A similar issue was discussed by Campbell [3], although his solution suggested approximation in the speech signal's feature space rather than a substitution at the phonetic representation.

The first polyglot crossover is at point B in Figure 1, after text processing but before prosody prediction. At this crossover point the English POS and syntactic role tags are mapped to the German tags. Both mappings are based on predefined mapping tables. This approach is different from that described by Pfister and Romsdorfer [8] where "inclusion grammars" give mutual connection between grammatical analysis results.

At point C, the crossover is after both text processing and prosody prediction but before the creation of the speech signal. The English phones are mapped to German at this crossover as described above. Pitch and duration values for the mapped English phones are obtained from the original English prosody predictions. POS tags and syntactic rules are not used by the synthesis component so are not relevant to crossover point C.

With these two crossover points, four different system configurations were defined, as shown in Table 1. For mixed-lingual texts, two 'hybrid' configurations were defined to process English/German text fragments: EEG/GGG and EGG/GGG. In all cases the same German voice is used for the waveform generation. For the EEG/GGG configuration, the text processing and the prosody of the fragments are specific to the native text language. In the EGG/GGG configuration only the text processing is language specific, with German prosody used on all the input. The English and German monolingual systems (EEE, GGG) were used as references for comparison in perceptual evaluations. Synthesis of mixed-lingual texts using the English voice was not investigated here, so crossovers from the German to English processing streams are not considered.

Table 1: Summary of the system configurations.

System Configuration	Text Analysis	Prosody Prediction	Voice
EEE	English	English	English
EEG	English	English	German
EGG	English	German	German
GGG	German	German	German

For mixed-lingual texts, the problem of language identification was eliminated by tagging the input texts with the appropriate language tag. At the language selection point, A (Figure 1), the input is divided into language fragments at the language tags and directed to the English or German text processing module as appropriate. For all German fragments, German prosody prediction is performed. For English fragments, processing then continues depending on the system configuration. In the EGG configuration, the phones, POS and syntactic role tags for English fragments are mapped at this point. Then the language fragments are re-assembled in their original order and processing continues with the German prosody and synthesizer modules. In the EEG configuration, the English fragments are processed by the English prosody module then the English phones are mapped to their German equivalents (the pitch and duration values are left intact) and the language fragments are re-assembled in their original order before processing continues with the synthesizer module.

3. Experimental setup

Perceptual experiments were run to assess how intelligibility, naturalness of prosody and acceptability vary depending on the polyglot system used and the amount of foreign inclusions in the input text. Four categories of input sentences were used: purely English sentences ("Eng"); purely German sentences ("Ger"); German sentences with English inclusions

of only a few words in length ("Inc"); and mixed-lingual sentences which contain full clauses of English and German ("Mix"). Each category consisted of 10 sentences. For example, one of the "Inc" sentences was: "*Am Trafalgar Square, nahe der Lord Nelson Statue, wird jedes Jahr ein riesiger Weihnachtsbaum errichtet.*". An example of the "Mix" sentences is "*Gravitation is not responsible for people falling in love – hat einmal Albert Einstein gesagt.*". Table 2 shows the average sentence length in words from each language in each text category.

Table 2: Summary of the German and English content in each input sentence category.

Name	Text Category	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

Table 3: Sentence and configuration combinations used to create experimental stimuli.

Cat.	System config	System explanation
Eng	**EEE**	Monolingual English TTS
	EEG	English text processing and prosody; German voice
	EGG	English text processing; German prosody and voice
	GGG	Monolingual German TTS
Ger	**GGG**	Monolingual German TTS
Inc	**EEG/GGG**	English parts: English text processing and prosody; German parts: German text processing and prosody German voice for both
	EGG/GGG	English parts: English text processing; German parts: German text processing; German prosody and voice for both
	GGG	Monolingual German TTS
Mix	**EEG/GGG**	English parts: English text processing and prosody; German parts: German text processing and prosody German voice for both
	EGG/GGG	English parts: English text processing; German parts: German text processing; German prosody and voice for both
	GGG	Monolingual German TTS

Table 3 summarises how the 4 input text categories were combined with the 4 system configurations to generate 11 different types of stimuli.

For the English sentences (Eng), all four system configurations were used to generate the utterances. The monolingual German TTS (GGG) tries to model how a native German person with no knowledge of English would read an English sentence and provides a lower bound for comparison of the performance of the polyglot system.

The German sentences (Ger) were only processed with the monolingual German TTS. These examples were used to obtain a baseline performance measure for the German monolingual system.

The mixed-lingual texts ("Inc", "Mix") were processed using hybrid systems EEG/GGG and EGG/GGG and the monolingual German TTS, which was used as a baseline for comparison of performance.

Each of the 10 sentences in each input category was produced by the configurations for that category, resulting in a total of 110 utterances for evaluation.

4. Evaluation

For perceptual evaluation, assessments were made by 26 subjects, 13 of whom were native speakers of English and 13 native speakers of German. All of the participants had at least a good knowledge of the other language.

The utterance set was split into two blocks of 55 utterances each, to keep one listening session below half an hour. The subjects were encouraged to take a 10–15 minute break between the two sessions. Prior to starting the experiments, the subjects were presented with 5 test utterances. These were selected to reflect the best and worst quality speech samples to familiarise the participants with the expected quality range of the utterances. Subjects were asked to assess each utterance for intelligibility, naturalness of prosody and acceptability of the utterances, using a mean opinion (MOS) score of 1–5 (1 denoting the worst and 5 the best).

To assess intelligibility, subjects were asked to score the utterance based on the amount of effort which they felt they were required to make in order to understand the gist of the message. Ratings were from completely unintelligible to perfectly intelligible.

For the naturalness of prosody, subjects were asked to assess the naturalness of the intonation, melody and rhythm of the utterance. To assist them, they were advised to indicate how human-like they found these components on a scale from fully machine-like to natural and human-like.

Finally to assess acceptability, subjects were asked how much they would accept the utterance as an answer from a machine in a voice response system from totally unacceptable to completely acceptable.

5. Discussion

Table 4 shows the results of the evaluation. Each result for intelligibility, naturalness and acceptability is shown separately for native speakers of English (Eng) and German (Ger). Each cell in the table contains the average of 130 values (10 different utterances, assessed by 13 native speakers).

Table 4: Results of perceptual evaluations (MOS on range 1–5, 1: worst, 5: best).

Setup		Intelligibility		Naturalness of Prosody		Acceptability	
Text	**System**	**Eng**	**Ger**	**Eng**	**Ger**	**Eng**	**Ger**
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

The primary purpose of the evaluation was to test two assumptions. Firstly, for longer foreign inclusions or complete texts in a foreign language, greater assimilation to the foreign language is required by the listeners in terms of prosody. That is, longer portions of foreign English speech were expected to sound more intelligible and natural when their prosodic structure was based on English rather than the German prosody of the primary voice. Secondly, for short inclusions of a few words the prosodic effect across languages was expected to be negligible so that the addition of English prosody would not cause significant differences in the perception in the synthesised speech quality.

The validity of the first assumption is illustrated by the results highlighted in Table 4, which compare systems to process English text with and without English prosody (EEG and EGG, respectively), on longer inclusions of English texts ("Mix") and complete English texts ("Eng"). For both cases in all the evaluation classes the systems with English prosody (EEG) were preferred (all mean differences in adjacent rows were statistically significant when tested by paired t-test with a 95% confidence interval).

In addition the second assumption is shown to be correct from the results highlighted for shorter inclusions ("Inc"). In this case there was no clear preference between the two systems (mean differences were not statistically significant). Since there is no significant difference, a single system which applies foreign prosody would be suitable for any length of foreign inclusion. This has the benefit that the system will not need to assess "how long is a piece of text".

The results also show a clear preference of both German and English native speakers for the use of language specific processing. As shown in Table 4, there is a big improvement in the evaluation scores for intelligibility and acceptability from the German only system (GGG) to the systems with English processing (all mean differences in adjacent rows from GGG → EGG were statistically significant). In terms of naturalness of prosody, since the EGG system used the same prosodic model as the German only system, unsurprisingly the trend was not so pronounced. Mean differences were not significant on full English texts ("Eng") for all subjects and on longer inclusions ("Mix") for German subjects. Overall this confirms the basic assumption that using a native linguistic model (pronunciation generation and syntactic analysis) is beneficial for mixed-lingual speech.

The performance of the German monolingual system (GGG) on purely German texts ("Ger") provides an upper bound on the expected performance of polyglot configurations on predominantly German texts ("Inc"). None of the results for EEG or EGG configurations on "Inc" stimuli exceeded this bound (the differences were statistically significant for all performance measures, except prosody performance assessed by native English speakers). It also confirms that assessments made on the mixed-lingual texts with longer inclusions ("Mix") were predominantly influenced by the performance on the English fragments, since none of the "Mix" results for EEG configurations exceeded the upper bound for performance on German fragments (the differences were all statistically significant, with the exception of German native speakers assessment of prosody).

Similarly, the performance of the English monolingual system (EEE) on purely English texts ("Eng") provides an expected upper bound for the performance of polyglot configurations on predominantly English texts. None of the results for EEG configurations on "Eng" stimuli exceeded this upper bound (all differences in performance between EEE and EEG configurations were statistically significant, with the exception of prosody assessment by native English speakers).

Comparing the performance of the monolingual systems on native texts, both English and German native speakers gave higher ranking to the intelligibility of the German system (differences were statistically significant). German native speakers ranked the prosody and acceptability of the English monolingual system higher than the English native speakers (the difference for intelligibility was not statistically significant). Similarly, for EEG configurations on "Eng" texts, German listeners again tended to give higher prosody and acceptability scores than the English subjects for the same setup. This suggests that the German listeners are more forgiving of errors heard in the other language. However, on "Mix" texts for EEG configuration, only the difference in prosody scores was significant between German and English subjects. For predominantly German texts ("Inc"), English native speakers gave higher scores in all categories for all system configurations, but the differences were not significant.

One factor that is not accounted for in these evaluations is the pleasantness of the synthetic voices. When questioned about their preference after the evaluations, both the English and German native speakers reported that they found the German voice more pleasant to listen to generally and in comparison with the US English female voice.

6. Conclusions

The aim of this study was to test the assumption that for synthesising mixed-lingual texts using a polyglot system with

a primary language, application of native prosody to the foreign inclusions becomes more important as the length of the inclusion increases. This assumption was tested through perceptual evaluations using an English-German polyglot system with German as the primary language. Systems with: no native processing; a native linguistic model; and with native linguistic and prosody models; were compared on mixed English-German texts with varying lengths of inclusion, and purely English texts. The systems were assessed in terms of intelligibility, naturalness of prosody and acceptability. These evaluations showed that for longer inclusions of, and fully, English text applying native prosody is preferable to using the primary language prosody for the foreign inclusions. For short inclusions the use of a native linguistic model is sufficient. However no degradation was observed when English prosody was applied to the English text in this case so if a system is expected to handle any length of inclusion a single system can be built which is independent of the text length. It was also found that listeners tend to be more forgiving of TTS errors in utterances in their non-native language.

7. Acknowledgements

The authors wish to thank for Stephen Gale for the design and implementation of the evaluation software tool and Dmitry Sityaev for his valuable advice on performing the perceptual evaluations.

8. References

[1] Traber, C. et al, "From Multilingual to Polyglot Speech Synthesis", *In Proc. Eurospeech'99*, pp. 835–838, Budapest, September, 1999.

[2] Tomokiyo, L.M., Black, A.W., Lenzo, K.A., "Foreign Accents in Synthetic Speech: Development and Evaluation", *In Proc. Interspeech'2005*, pp. 1469–1472, Lisbon, September, 2005.

[3] Campbell, N. "Talking Foreign. Concatenative Speech Synthesis and the Language Barrier", *In Proc. Eurospeech'2001*, pp. 337–340, Aalborg, September, 2001.

[4] Badino, L., Barolo, C., Quazza, S., "Language Independent Phoneme Mapping For Foreign TTS", *In Proc. Fifth ISCA ITRW on Speech Synthesis*, pp. 217–218, Pittsburgh, June, 2004.

[5] Mashimo, M. et al, "Evaluation of Cross-Language Voice Conversion Based on GMM and STRAIGHT", *In Proc. Eurospeech'2001*, pp. 361–364, Aalborg, September, 2001.

[6] Latorre, J., Iwano, K., Furui, S., "Polyglot Synthesis Using a Mixture of Monolingual Corpora", *In Proc. 2005 ICASSP 2005*, pp. I-1 – I-4, Philadelphia, March, 2005.

[7] Black, A. et al, "Multilingual Text-to-Speech Synthesis", *In Proc. ICASSP 2004*, pp. III-761 – III-764, Montreal, May, 2004.

[8] Pfister, B., Romsdorfer, H., "Mixed-Lingual Text Analysis for Polyglot TTS Synthesis", *In Proc. Eurospeech'03*, pp. 2037–2040, Geneva, September, 2003.