Building Text-to-Speech Systems for Resource Poor Languages

Nur Hana Samsudin and Mark Lee
{n.h.samsudin and m.g.lee }@cs.bham.ac.uk

Introduction

The focus of this research is to develop a method for building Text to Speech Systems for resource poor languages by using data from other languages to fine tune a general template polyglot TTS architecture.

Our method involves three main componants: language clustering, phoneme mappings and prosody modelling. As a proof of concept, four TTS have been implemented for English, Spanish, Malay and Iban as follows.

1. English TTS from German data
2. Spanish TTS from Indonesian data
3. Malay TTSs from English and Afrikaans data
4. Iban TTSs from Malay, Indonesian and Spanish data

The Components

The three components in brief:

- Language clustering: clustering the language features into the individual description of the target language
- Global phoneme mapping: defining the global phoneme and identifying two template of grapheme to phoneme conversion.
- Prosody modelling: determine the template of the prosody based on the INTSINT modelling

To ease the adaptation from one language to another, a phoneme substitution matrix has been introduced. This matrix provides the possible phoneme substitution when the resource language do not have the target language's phoneme.

Resources

This research uses the multilingual data provided by different organisation and researcher:

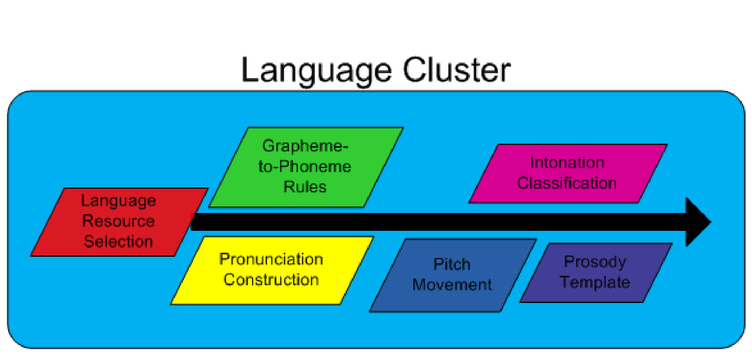
- MBROLA - database by Faculté Polytechnique de Mons, Belgium for language and voice data
- The International Phonetic Alphabet - by the International Phonetic Association for phoneme set and notation
- SAMPA - by J. Wells as a computer readable phonetic alphabets
- INTSINT - International Symbol of Intonation was created and improvises by Daniel Hirst of Université de Provence.

Acknowledgements

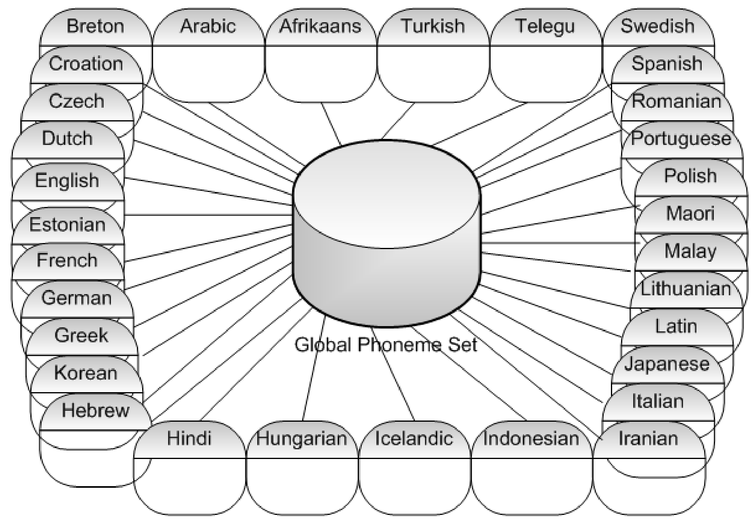
The authors acknowledge the contributions of the researchers from Sarawak Language Technology (SaLT) Research Group at Universiti Malaysia Sarawak (UniMAS). Special appreciation goes to Sarah Flora Samson Juan for preparing the Iban phoneme set.

The Language Adaptation

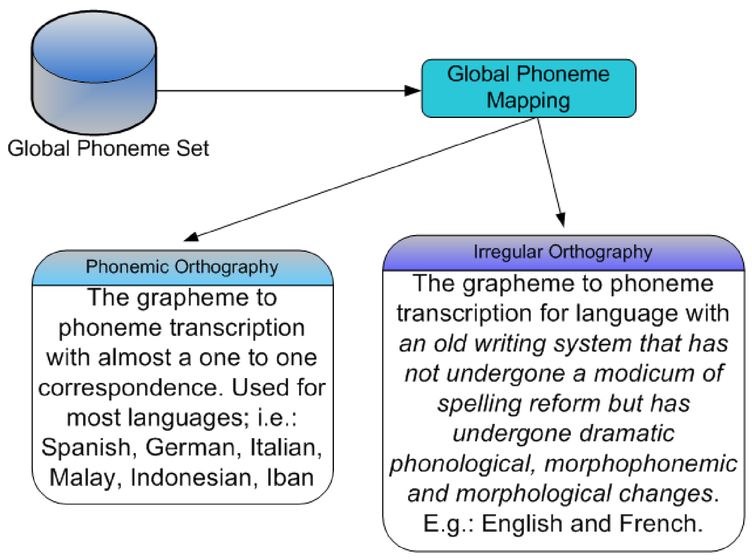
Language Clusters Processes



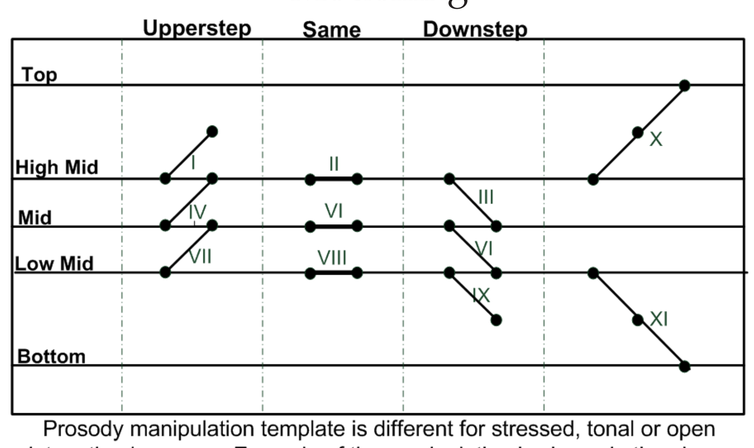
Creating Global Phoneme Set



Identifying Grapheme to Phoneme Mapping



Prosody Modelling



Prosody manipulation template is different for stressed, tonal or open intonation language. Example of the manipulation is shown in the above lines. Intonation with the lines I, IV and VII holds the previous pitch's value or INTSINT value; **M**edium; and at the end syllable, the pitch holds the **U**pperstep value. For the lines II, V and VIII, no changes of initial and final pitch of the syllable occur. For the lines III, VI and IX, the final values will hold the **D**ownstep value. While for the lines X and XI, the calculation of **U**pperstep and **D**ownstep will be calculated twice respectively.

Results

Repondents in general can understand most of the synthesised speech. Spanish respondents mostly preferred the TTS using the monolingual TTS, while there were similarities in the preferences of English and Malay respondents. Individual comments on different TTS adaptation were obtained.

1. English
   Some German phoneme is not as strongly produced as English. This is a voice coder issue.
2. Spanish
   Spanish polyglot is substantially poorer than the monolingual, but was perceived correctly.
3. Malay
   English resources are preferable to Malay. There was a noticeable accent from English resources but this did not restrict understanding.
4. Iban
   Indonesian resource are the most preferred resource. Iban using Spanish resource are generated using Spanish prosody and thus respondents acknowledge the slightly faster pace than it should be.

More detailed results are presented in the full paper.

Conclusion

This paper presented an approach to adapting available resources to create TTSs for resource poor languages. To evaluate our approach, TTSs for two resource rich languages were implemented, namely English and Spanish. These polyglot TTSs are compared to a monolingual TTS. TTSs for two resource poor languages, Malay and Iban, were also constructed. The Malay polyglot TTS was developed using English and Afrikaans data and the Iban polyglot TTS was developed using Malay, Indonesian and Spanish data.

The approach has demonstrated that it is capable of producing acceptable polyglot speech synthesis from different language data. Although the output is not as good as a monolingual TTS for English, Spanish and Malay, the polyglot synthesised speech is acceptable to native speakers.

The respondents prefer Iban TTS generated from Indonesian data (which match the Iban language features) although the TTS using Malay resource is mimicking the native speakers' tone. Spanish resource are also preferred although respondents felt the synthesised speech has a faster pace than it should.

Future research will further investigate the factors behind the relative effectiveness of different related language resources.