

Loquendo Mixed-Language Support

The fact that the quality of Text-To-Speech (TTS) rendering has greatly improved in recent years is well known. The new generation of speech synthesis engines is based on a large database of recorded speech; we call this technique *Unit Selection* and we were among the first companies to deploy it in a TTS product. (See www.loquendo.com for more details on the Unit Selection speech synthesis technique.) The main goal of the previous generation of TTS systems was to reach intelligibility, but the produced speech sounded still unnatural, robotic-like and so unfriendly, even if it was almost perfectly intelligible.

The quality today is good, but obviously it can be improved further. A TTS challenge that will be analyzed in this paper is how to render a text that mixes more than one language by TTS. We call this *mixed-language* capability and it is a new feature recently added to the Loquendo TTS (vers. 6.2). Is this feature useful? I'd say yes, think of texts coming from different sources in unpredictable languages (e.g. Internet), or e-mails that may contain puzzles of attached pieces written in a mixture of languages, or advertising slogans, movie titles, reported news, proper names, song titles, etc.

How to address mixed-language? If you simply try to render a mixed-language text with a mono-lingual synthesis engine, the result will be at best funny, at worst almost incomprehensible. Rules of phonology, coarticulation, morphology, and syntax will clash together to produce an awful result, not useful at all for a real life speech application.

The first idea is to choose a bi-lingual, or even multi-lingual, let us say, a polyglot speaker talent and record her/him speaking in different languages. Fine, but there is still a technical issue, which is the language of texts to be rendered by TTS. You need a *language guesser*! It is a software that is able to guess the language of a piece of text. Loquendo has implemented it!

The *Loquendo Language Guesser* is trained on a large amount of texts in different languages to improve its discriminative power. It actually covers all the languages deployed by Loquendo TTS. The prediction accuracy is higher for longer text portions and can be improved by reducing the number of languages in alternative.

Let us start with an example, if the sentence to be spoken is: "Hello Mme. Françoise Dupont can I help you?" the Language Guesser has to identify that: "Mme. Françoise Dupont" is to be spoken in French. So that the result of the Language Guesser may be the following: "Hello **lang=French** Mme. Françoise Dupont **lang** can I help you?", that actually is the format in which you can direct Loquendo TTS if you want to tag different languages by yourself.

It is worth noticing that to select a language by the Language Guesser allows the synthesis engine to change the language knowledge that is mandatory for TTS. The first step is to apply a language dependent text processing, for instance in the previous example "Mme." has to be expanded into the French word "madame" (corresponding to "Mrs" or "madam" in English). Then the word has to be transformed into a string of phonemes in French, therefore "madame" will become /mad'am/ as it is spoken in French, instead of /'mædəm/ as it is spoken in English. The phonemes are the units that compose a spoken language, and in the examples are represented in a simplified form. Actually the phoneme-sets for different languages are different.

Is the mixed-language issue solved? Not at all, even if the Language Guesser helps to select the correct transcriber in the speech engine, it is very difficult to find a good bi-lingual speaker. To find a penta-lingual, or an epta-lingual speaker is a mission impossible! In the reality we have successfully found Castilian and Catalan speakers, for instance our Carmen, Spanish Castilian female voice, and Montserrat, Catalan female voice, are recorded by the same speaker, and also the male voices Jordi, Catalan, and Jorge, Castilian Spanish, too. This is the exception that confirms the rule!

Another possible option is to use the Language Guesser to select the language and then to switch to a different voice of the guessed language. This is possible and in some contexts it is a good solution, but in

general it is very annoying to have voice changes too often. In the example above, you should change the voice in the middle of a sentence, that is not very nice.

The final idea was to try to map the phonemes of one language into another, we call this technique Phoneme Mapping. Let us try to describe it: the idea is that if a text written in language *lang1*, which includes in it a piece written in language *lang2*, we need to first identify the right piece written in *lang2*, by using the Language Guesser, then to transform it to the phonemes of the *lang2* language, as a native speaker will do for reading it, and finally to map the phoneme string into the container language *lang1* phoneme-set.

Let us look at the previous example:

Hello Vang=French Mme Francois Dupont Vang can I help you?

The first step is to transcribe each piece of text with a different transcriber according to the tagging done by the Language Guesser. The result is the following where the transcribed French words are marked:

/həl'əʊ/ /mad'am/ /fʁɑ̃sw'az/ /dyp'ɔ̃/ /kæn/ /aɪ/ /h'elp^h/ /ju/

The next step is to map the French phonemes into English ones. The result is the following:

/həl'əʊ/ /ma:d'a:m/ /fɹɑ:nsw'a:z/ /dup^h'ɔ:n/ /kæn/ /aɪ/ /h'elp̃/ /ju/

The last step is a re-processing of the English transcription to perform an allophonic substitution.

Pros&Cons: The result is that the whole text is rendered by the same voice and you are not required to change the voice in the middle of a sentence to take into account the inserted language. The *Phoneme Mapping* is a kind of an approximated pronunciation. It will sound like a well trained English speaker reading French. At a first sight this approximation seems to be a drawback; however, it is actually a quite good result, because if you mix an English text into a correctly pronounced French piece, you will run into a range of difficult problems.

In fact, a speaker having to pronounce foreign words included in a text written predominantly in her/his own language will be inclined to pronounce these words in a manner that may differ - also significantly - from the correct pronunciation of the same words when included in a complete text in the corresponding foreign language, for instance coarticulation into the two languages is different. The approximation of this kind of pronunciation is especially due to the speaker choice of maintaining his native-tongue phonological system, but also to co-articulation, economy of effort, and to psychosocial factors. Even a trained speaker will do the same approximation to avoid the worst tongue contortions.

References:

If you are interested in going more deeply into how a speech synthesis engine is implemented or finding out more about mixed-language capabilities, you are invited to read some additional papers, such as the following:

Silvia Quazza, Laura Donetti, Loreta Moisa, Pier Luigi Salza, "**ACTOR®: A Multilingual Unit-Selection Speech Synthesis System**", *Proc. of 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Atholl, Scotland, 2001.
[\[http://www.loquendo.com/en/brochure/art_TTS_2001.pdf\]](http://www.loquendo.com/en/brochure/art_TTS_2001.pdf)

Badino Leonardo, Barolo Claudia, Quazza Silvia, "**A General Approach to TTS Reading of Mixed-Language Texts**", *Proc. of 5th ISCA Tutorial and Research Workshop on Speech Synthesis*, Pittsburgh, PA, 2004. [\[tp://www.loquendo.com/en/brochure/art_ml_2004.pdf\]](http://www.loquendo.com/en/brochure/art_ml_2004.pdf)