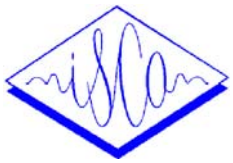
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

A General Approach to TTS Reading of Mixed-Language Texts

Badino Leonardo, Barolo Claudia, Quazza Silvia

Loquendo S.p.A., Voice Technologies
{leonardo.badino, claudia.barolo,
silvia.quazza}@loquendo.com

Abstract

The paper presents the Loquendo TTS approach to mixed-language speech synthesis, offering a range of options to face the various situations where texts may occur in different languages or embedding foreign phrases. The most challenging target is to make a monolingual TTS voice read a foreign language text. The adopted Foreign Pronunciation Strategy here discussed allows mixing phonetic transcriptions of different languages, relying on a Phoneme Mapping algorithm making foreign phoneme sequences pronounceable by monolingual voices. The algorithm extends previous solutions, obtaining a plausible approximated pronunciation. The method is efficient, language independent, entirely phonetics-based and it enables any Loquendo TTS voice to speak all the languages provided by the system.

10.21437/Interspeech.2004-312

1. Introduction

A text-to-speech system based on unit selection concatenative synthesis, like Loquendo TTS [1], relies on a speech database of pre-recorded sentences pronounced by mother-tongue speakers. The voice database is single-language in that all the sentences are written and pronounced in the speaker native language, so that the acoustic units available for concatenation belong to a single prosodic-phonological system. Moreover, all text-analysis functions in a TTS are language-specific, admitting as exceptions a few foreign words transcribed in a pronunciation lexicon. Basically, traditional systems are conceived to read monolingual texts. Multilingual texts can be correctly read by changing the voice at every language change, what can be unfeasible for truly *mixed*-language texts, where changes occur frequently and are embedded in sentences and phrases. Real applications would require a more flexible behavior to handle a variety of situations, e.g. texts coming from different sources in unpredictable language (e.g. internet), e-mails or office documents written mainly in a language and partially in a second language (typically English), messages including foreign names or phrases (e.g. film titles) for an information service, etc. In some of these cases the optimal solution would be to have the same TTS voice reading the whole mixed-language text. This solution has been pursued in recent years by adopting essentially two different approaches. On the one hand [2], attempts were made of producing multi lingual vocalic databases by resorting to bilingual or multi lingual speakers. Unfortunately this *polyglot* approach is based on assumptions (essentially, the availability of a multi-lingual speaker) that are seldom true.
Another approach consists in applying an automatic phonetic transcriber for the foreign language and then *mapping* the obtained transcription onto the phonemes of the native language of the voice, in order to access its acoustic units. The idea was introduced in [3] where a Japanese TTS voice could pronounce an English text thanks to a correspondence table between English and Japanese phonemes and to a method for finding acoustically suitable units. The method was extended and refined in [4], where a fine allophonic labeling of the speech data was at the basis of the mapping.
While the first approach realizes a “perfect pronunciation”, the second brings an “approximate pronunciation”. This doesn’t mean that the first method is better than the second: looking at many real cases, the approximate approach, may fit better to reality. In fact, a speaker having to pronounce foreign words included in a text written predominantly in his or her own language will be generally 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. The approximation of this kind of pronunciation is especially due to the speaker choice of maintaining his native-tongue phonological system. This choice is due to co-articulation, economy of effort and also to psychosocial factors, as adopting the correct pronunciation may be regarded as an undue sophistication and, as such, rejected in common usage.
As discussed in the following, in the implementation of an overall strategy for Loquendo TTS to tackle mixed-languages texts, we applied both the described approaches to foreign text pronunciation. According to the first one, we realized two bilingual acoustic dictionaries, recording two Castilian-Catalan mother-tongue speakers. Following the second approach, we came to a quite general and language-independent solution, aiming to obtain a plausible phoneme mapping between any pair of languages on the basis of a general concept of phonetic similarity.

2. A Strategy for mixed-language TTS

The Loquendo TTS system is conceived according to a multi-lingual modular architecture. A language-independent engine performs text-to-speech conversion by applying language-specific functions and knowledge bases, available in separate Dynamic Link Libraries. Such design allows switching between languages on the fly and even mixing functions from different DLL' s. A Language Guessing module can guess the language of the text. The Guesser is statistically based and is trained on large word lists belonging to the languages supported by the system. The prediction accuracy is higher for longer text portions and can be improved by reducing the number of alternatives. A wide range of solutions is available

to manage multi-language texts. The following are typical cases:

- Guess the language of a whole text and switch to the suitable voice. Loquendo TTS provides voices in a number of languages (Italian, French, German, Greek, Dutch, Swedish, Chinese, Catalan, several varieties of Spanish, English and Portuguese).
- Guess the language of each paragraph (e.g. in an e-mail reader) and switch to the suitable language while keeping the same voice.
- Switch to the language specified via a control tag in a marked-up text (e.g. for embedded foreign phrases in an information service)

The last two cases can be dealt by resorting to bi-lingual voices (at present available for the Spanish/Catalan pair) or by applying the Foreign Pronunciation strategy (see Fig.1). The latter amounts to flexibly mixing the phonetic transcription functions of two languages. In this case, a voice native of language L1 is forced to pronounce portions of text in language L2.

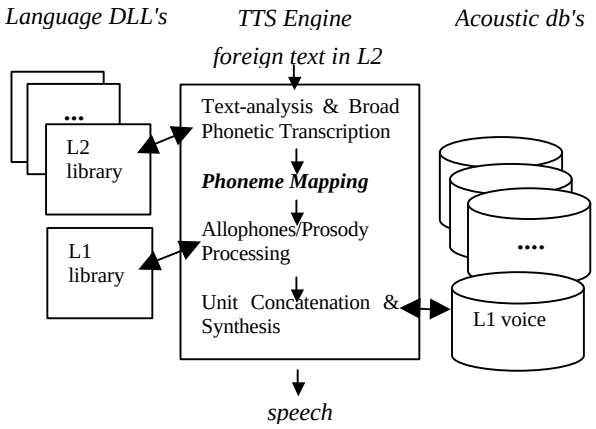


Figure 1. *Foreign pronunciation flow in the TTS architecture*

The text-analysis functions in the L2 DLL are invoked on the foreign text portion in order to obtain its broad phonetic transcription. Then the L2 phoneme stream is mapped onto L1 phonemes by the Phoneme Mapping Algorithm (described in Paragraph 3) and inserted in the L1 phonetic transcription of the whole text. From this point on, the L1 DLL functions are invoked, applying allophonic and prosodic rules and concatenating speech units from the L1 acoustic database. The application of L1 phonetics and prosody ensures maximum coherence with the contents of the L1 acoustic database. As discussed in the following (see Paragraph 5), this amounts to adopting the point of view of native L1 speakers and listeners, who may know the correct pronunciation of foreign words but adapt them to their own phonological system.

3. Phoneme Mapping

The key point in the Foreign Pronunciation Strategy is the Phoneme Mapping algorithm. While the flexible Loquendo TTS architecture can provide phonetic transcriptions where each word is transcribed according to its language, a further step is required in order to obtain their pronunciation by a single-language voice. Phonemes that do not belong to the native phonological system of the voice must be replaced by the most similar sounds available in the voice acoustic database. To this end we have implemented a quite general and language-independent algorithm intended to convert a string of L2 phonemes into the closest L1 phoneme string. The algorithm centers around a Similarity Function, computing a similarity score between two phonemes depending on their phonetic-articulatory features.

3.1. The Phoneme Mapping Algorithm

The Phoneme Mapping algorithm is implemented in the TTS engine and acts independently of any language-specific knowledge. Its inputs are the string of L2 phonemes to be converted and the phoneme inventory of the target language L1. The input string is scanned left-to-right focusing on a single phoneme at a time. The focused L2 phoneme is compared with every L1 phoneme in the L1 inventory, obtaining scores by the Similarity Function (described below). The L1 phoneme with the highest score is selected and appended to the output string, provided that the score is above a predefined threshold. In case no phoneme is found with a suitable score, the output phoneme is null. This may happen for example for an English /h/, for which no similar sound can be found in an Italian or French voice. In other cases, the input L2 phoneme is best rendered by a sequence of two L1 phonemes. This may be true for complex phonemes, such as nasalized or rothacized vowels, diphthongs and affricates. If the target language lacks the corresponding phoneme, the algorithm would obtain a similar sound by composition. For example, in a French to English mapping, a nasalized vowel would be mapped onto the corresponding simple vowel and a nasal consonant would be added. For affricates and diphthongs, a double search is attempted, comparing them both with single L1 phonemes and with phoneme pairs. For instance, an English diphthong would be mapped onto the corresponding German diphthong, if it exists, otherwise it would yield a vowel pair.

3.2. The Phonetic Similarity function

A language-independent Phoneme Mapping is feasible only if the similarity between two phonemes can be judged without referring either to the phonological systems to which they belong, or to any other language-specific knowledge. A quite general classification of phonemes is necessary, such as the one defined by Articulatory Phonetics. Our working hypothesis was that two phonemes are perceived as similar when they have similar phonetic-articulatory features. This was clearly a strong assumption, overlooking finer and language-dependent aspects of speech perception, as well as pragmatic/cultural factors that may affect the pronunciation of foreign languages. Nevertheless, it provided a useful basis for the implementation of a Phonetic Similarity function, yielding satisfying results in a computationally efficient way.
The idea was to represent each phoneme as a vector of articulatory features, according to the concepts of classical phonetics [5]. The Phonetic Similarity function would compare two vectors and compute a score depending on the

distance between the vector values. A metrics was defined to compare feature values, while features themselves were assigned different weights in the score computation, according to their influence on the perception of similarity.

3.2.1. Feature vectors and comparison metrics

We defined vowel vectors as composed of “non-binary” categories specifying their position in the vowel quadrilateral [5], plus some additional binary properties (nasalized/non-nasalized, rhotacized/non-rhotacized, stressed/unstressed, etc.). For diphthongs, the position in the quadrilateral is specified for both their component vowels. Vectors describing consonants are composed of “non-binary” features referring to manner (i.e. nasal, fricative, approximant, affricate) and place of articulation (i.e. dental, alveolar, retroflex) plus some binary features (aspirated/non-aspirated, syllabic/non-syllabic, released/unreleased, etc…).
The perception of similarity may be affected to different degrees by the different features. For instance, in the vowels comparison the rounded/non-rounded feature seems to be more discriminating than the stressed/unstressed one. Besides, the different values of a non-binary feature can be placed on a scale of perceptual distance (e.g. post-alveolar is closer to retroflex rather than to alveolar). The challenge was to define weights for the features and distances for their values in such a way that the resulting similarity score be in accordance with perception. To this end we applied an iterative process. As a first step we implemented a rough mapping module in which all the features had the same importance and their values were equivalent. Then we performed an informal perceptual test where mother-tongue subjects were asked to evaluate the intelligibility and plausibility of foreign words synthesized with the various Loquendo voices via the mapping module. The languages involved in the test were Catalan, Chinese, English, French, German, Greek, Italian, Spanish and Swedish. On the basis of test results, we re-defined weights and distances. This process was iterated until perceptual tests gave satisfying results for all the language pairs.

3.2.2. Exceptions

The similarity function handles a small number of exceptions to the general assumption that phonemes with similar phonetics features are perceived as similar, independently of the language. A special case is that of the pronunciation of the letter “r”, realized in different languages with phonetically very distant phonemes, which nevertheless are often perceived as similar (e.g. the German fricative-uvular-voiced /ʁ/ vs. the Italian trill-alveolar-voiced /r/). In a few cases we actually found a different perception of similarity by listeners of different mother tongue, but we were able to maintain to our mapping its language independence, by forcing a compromise choice ensuring intelligibility. This was the case for the English fricative consonant /ð/, sounding like a dental-plosive to an Italian listener and like a fricative-alveolar to a French listener. We decided to map /ð/ onto the dental-plosive (when available for the target voice), ensuring an intelligible English pronunciation for all the TTS voices.

4. Results

An example of application of the Foreign Pronunciation Strategy is illustrated in Figure2, where a mixed-language text, embedding French words into an English sentence, is converted into a sequence of English phonemes. The foreign pronunciation is in this case activated by the control tag "*lang=*" in the input text.

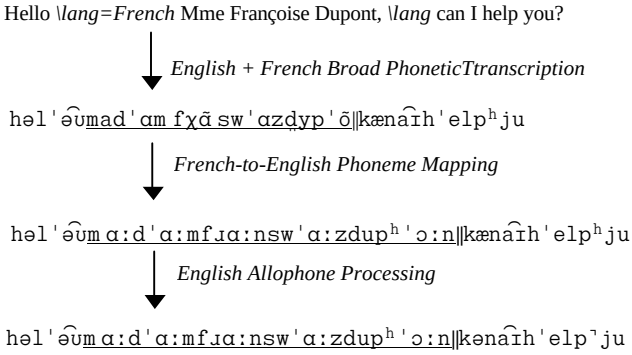


Figure 2. *Phonetic transcription of a mixed-language text*

The example shows that foreign text may be actually embedded inside sentences. Here the foreign portion is explicitly marked with control tags. Alternatively, the Language Guesser may automatically detect the foreign phrase. While the English DLL processes English words, the French DLL functions are applied on the French text portion, expanding the abbreviation "Mme" into the word "madame", converting graphemes into phonemes and placing word stress. The Phoneme Mapping algorithm is then applied, mapping each French phoneme onto a phoneme from the English inventory. In the example, close mappings are found in some cases, such as [y] : [u], differing for the front/back feature, or [p] : [pʰ], differing only for aspiration. Other cases are less obvious. The uvular unvoiced fricative [χ] is mapped onto the alveolar approximant [ɹ], as both may be considered allophones of /r/. The two nasalized vowels [ɑ̃], [õ] are mapped onto the vowels [ɑː], [ɔː], respectively, while the nasalization is rendered by inserting a following [n]. The last step towards a narrow English transcription is the application of contextual allophonic rules, which in the example replace [æ] with [ə] and [pʰ] with the unreleased plosive [pˀ].
The foreign pronunciation by Loquendo TTS voices can be directly experienced and tested on the Loquendo web site (http://www.loquendo.com). As it can be heard, the effect is plausible and captures the typical behavior of foreign speakers, resulting almost in a caricature of the various foreign accents.

5. Discussion

The quite general approach adopted in the definition of the Phoneme Similarity function, seems to obtain satisfying results, in most cases yielding the same phoneme mapping that would have been manually selected. On the other hand, its generality is obviously a great advantage in the development of a multi-lingual TTS, if compared with table look-up approaches [3] that would require creation and

updating of tables for each considered language pair. The easy extension of the algorithm to new languages has been proved by its application to Portuguese and Dutch, which were not among the languages on which it had been tuned. This does not entitle us to claim that the approach is universal. It turns out, expectedly, that better results are obtained for languages belonging to the same linguistic family, while the mapping is not convincing when tonal languages are involved (Chinese and Swedish, in our language set), as tones are not represented in feature vectors.

For non-tonal European languages, we might say that our working hypothesis basing the perceptual similarity on articulatory features has been substantially confirmed, with few exceptions (see Paragraph 3.2.2). What may be advisable is an extension of the only-phonetic definition of similarity that would include some cultural aspects related to grapheme representation and language evolution. An interesting case is that of the /r/, mentioned above (see Paragraphs 3.2.2 and 4). The considered European languages read the 'r' grapheme in very distant manners, which nevertheless are generally perceived as /r/'s. As an extreme case, the 'r' occurring in a word like "four" ([fɔ:]) is not pronounced at all by a British English speaker, but it is 'perceived' by a foreign speaker, who for example would map [fɔ:] into [fɔ:r], and re-appears as a *linking 'r'* also in British if the word is followed by a vowel.

The Loquendo Foreign Pronunciation Strategy based on Phoneme Mapping yields approximate pronunciations, basically correct but carrying a strong foreign accent, what is perfectly acceptable and even desirable for listeners sharing the same native language of the voice. Its intended application is for reading small portions of foreign text embedded in a text written predominantly in the native language of the TTS voice, where a switch in the prosodic and phonological system would sound unnatural and would hinder intelligibility. Phoneme Mapping may substitute pronunciation lexicons, at least when foreign words can be marked in the texts or guessed by the Language Guesser. Alternatively, it can help creating such lexicons off-line.

In our strategy, the exact point where to switch between languages is crucial. If we represent the phonetic transcription process as ideally performed in three steps, a first word-level broad transcription, as it would result from dictionary look-up, followed by the application of sentence-level phonological rules (e.g. assimilation, *liaison*, de-accentuation, etc.) and finally by finer co-articulation allophone modifications, we may argue that the third step pertains to the actual speech performance of the speaker, while the first two pertain to its linguistic competence. In this sense, the switch to the native language of the voice should occur just before the third step, when the broad transcription is converted to a narrow one. This choice is debatable and partially contrasts with what argued in [3,4]. In those works, a fine allophonic labeling of a Japanese database allowed to map the English /l/ and /r/ onto different allophones of the same Japanese phoneme. It is certainly true that in principle (and in many real cases), the availability of detailed L1 allophones could provide a mapping closer to the original L2 pronunciation. But such fine mimicking of a foreign phonetics may have some drawbacks. For example, the English voiceless plosives show two allophonic variants, the more common aspirated allophone and the non-aspirated one. When mapping a French or Italian non-aspirated plosive, the closest English allophone would be the non-aspirated one. In the example in Figure 2, the French [p] in the word [d̪ypˈõ] could have been mapped on the English [p] rather than on [pʰ]. What we argue is that such closer mapping might sound less plausible to an English listener, who may perceive the non-aspirated plosive as its voiced counterpart, if occurring in an unexpected context. In fact, the English allophone [p] is produced only after an [s], while in most other contexts the /p/ would be aspirated. In the intervocalic position of our example, the aspiration would be the main cue to distinguish /p/ and /d/, as the simple presence or absence of voicing is not relevant to an English hear [6]. Besides, no long units will be found in the speech database to cover a rare allophone in an improper context. This would increase the risk of discontinuities in unit concatenation, as the voice database is specifically designed to cover the most frequent phoneme sequences in the target language. When synthesizing foreign words, new phonetic contexts may arise that are not found in the database, what may cause a poorer voice quality in foreign pronunciation. This may be considered an intrinsic limitation of the Phonetic Mapping approach as applied to the unit-selection synthesis technique. When the phonetic distance between languages is great and the required voice quality is high, a compromise between the *polyglot* and the *mapping* approaches (see Introduction) may be in order. The voice database could be augmented with speech material specifically designed to cover phoneme sequences arising in Foreign Pronunciation or even foreign phonemes difficult to map on the voice phonological system. This is what we are currently implementing for some of the Loquendo voices. Due to the practical relevance of English, voices are being enriched in order to obtain a smoother foreign pronunciation at least for this language.

6. Conclusion

An optimal reading of mixed-language texts requires a variety of functionalities by a TTS system, including the capability of reading foreign text with a monolingual voice. The Foreign Pronunciation strategy here described follows the approach initiated in [3,4]. The original idea of mapping foreign phonemes onto the phonological system of the voice in order to access its vocal database has been developed into a general and efficient language-independent algorithm.

7. References

[1] Quazza S., Donetti L., Moisa L, Salza P.L., "ACTOR: a Multilingual Unit-Selection Speech Synthesis System", *4th ITRW on Speech Synthesis*, Perthshire Scotland, 2001

[2] Traber, C. et al.: "From multilingual to polyglot speech synthesis", *Proceedings of EUROSPEECH '99*, Budapest, 1999

[3] Campbell, W.N.,"Foreign-language speech synthesis", *Proc.ESCA ETRW on Speech Synthesis*, Jenolan Caves, Australia, 1998.

[4] Campbell, W.N., "Talking Foreign. Concatenative Speech Synthesis and Language Barrier", *Proceedings of EUROSPEECH 2001*, Scandinavia, 2001

[5] "Handbook of the International Phonetic Association.". Cambridge University Press, 1999.

[6] Roach, P., "English Phonetics and Phonology", Cambridge University Press, 1983