Language Independent Phoneme Mapping For Foreign TTS

Leonardo Badino, Claudia Barolo, Silvia Quazza

Loquendo, Vocal Technology and Services
Turin, Italy
{leonardo.badino,claudia.barolo,silvia.quazza}@loquendo.com

ABSTRACT

This note describes a tentative solution to the problem of mixed language texts in TTS applications. The multi-language modular architecture of Loquendo TTS has been exploited to provide a range of user options, allowing to guess the language of paragraphs/phrase/words and to switch between voices in different languages, or between foreign accents of the same voice. This note focuses on a Phoneme Mapping algorithm enabling any TTS voice to speak all the languages provided by the system. The approach is quite general and language independent, entirely phonetics-based. The obtained foreign pronunciation is by definition approximated but plausible and suitable to reading foreign words or phrases embedded in a text.

1. INTRODUCTION

Text-to-speech conversion is intrinsically language-dependent, from text formatting down to sound production. Nevertheless, real applications are more and more facing TTS systems with multi-lingual texts. A flexible behaviour is required. In some cases (e.g. texts from the web), once the language of the text has been detected, the system could simply switch to the suitable voice. In other cases, the voice should preferably remain the same but the pronunciation should be adapted to the foreign language. This is the case for foreign words or phrases embedded in a text, traditionally approached through pronunciation lexicons, but also for situations where a second language (typically English) is occasionally used for technical or international communication (e.g. e-mails in an office environment). The TTS system by Loquendo [1] provides a range of solutions for mixed-language synthesis, based on its multi-lingual modular architecture. A Language Guesser can be invoked to detect the language of the text, on a paragraph or phrase or word basis. A number of voices in different languages (Italian, French, German, Greek, Swedish, Chinese, Catalan, several varieties of Spanish, English and Portuguese) can be selected, depending on the language. For the Spanish-Catalan language pair, two (male and female) bilingual voices are available. A similar solution cannot be extended to every language pair. In order to allow any Loquendo voice to plausibly pronounce foreign words and phrases, a general solution has been envisaged, as described in the following.

2. FOREIGN LANGUAGE TTS

Thanks to the modular structure of Loquendo TTS, it is perfectly feasible to invoke different language libraries on different text portions, obtaining a phonetic transcription where each word is represented according to its language. The point is that such a transcription, mixing phonemes belonging to different phonological systems, cannot be pronounced by a single voice. In a unit-selection TTS, sounds are synthesized by extracting them from a speech database, typically recorded by a professional speaker in his native tongue. The database would generally contain phonemes of a single language. Consequently, foreign phonemes should be realized with some approximation by choosing the most *similar* phonemes in the database. This approach was first applied in [2] for mapping English onto Japanese and vice versa, basing on the strong assumption that two phonemes can be judged similar when they have similar phonetic-articulatory features. Such simple hypothesis, overlooking finer and language-dependent aspects of perception, turned out to be very handy for our purposes, allowing a computationally efficient access to similar speech signals and providing the basis for a quite general and language-independent phoneme-mapping algorithm. The resulting foreign language synthesis simulates the behaviour of a speaker who knows the correct pronunciation of the foreign word but does not switch to the foreign phonological and prosodic system, due to co-articulation reasons and economy of effort, or lack of fluency in the foreign language. This kind of approximated pronunciation is the most suitable to contexts where speaker and listener share the same mother tongue (e.g. reading foreign titles in a movie information service), although in real situations the adopted pronunciation would probably be affected to some extent by the grapheme-to-phoneme rules of the mother tongue and by other pragmatic factors.

2.1. The Phonetic Similarity Function

The core of the mapping algorithm is a function computing a similarity score between two phonemes. Its implementation required the following steps.

- Representing each phoneme as a vector of phonetic-articulatory features [3].
- Defining the weight of each articulatory feature in the similarity estimate.
- Computing the degree of affinity between the values of "non-binary" articulatory features.

We defined vowel vectors as composed of "non-binary" categories specifying their position in the vowel quadrilateral [3], 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 Similarity Function would estimate the similarity between two phonemes by comparing their feature vectors.

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. 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.

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.

2.2. The Phoneme Mapping algorithm

The foreign transcription algorithm is implemented in the TTS engine and invoked when a text portion is written in a language (L2) different from that of the active voice (L1). In that case, the engine first calls the text-to-broad-phonetic-transcription functions of the L2 library, then it applies the Phoneme Mapping function converting the L2 transcription into an L1 transcription, and finally it reverts to the L1 library functions to perform broad-to-narrow phonetic transcription and phoneme-to-sound conversion.

The Phoneme Mapping function receives as parameters the L1 and L2 phoneme inventories together with the string of L2 phonemes to be converted. Every L2 phoneme in the input string is compared with each L1 phoneme in the L1 inventory, obtaining scores by the Similarity Function. The L1 phoneme with the highest score is selected, provided that the score is above a predefined threshold, otherwise the output phoneme is null (e.g. an English /h/ is skipped in a Englih-to-Italian mapping). In some cases, the input L2 phoneme is best rendered by a sequence of two L1 phonemes. For example, for an L2 nasalized or rothacized vowel, if the closest L1 phoneme is a simple vowel, the Phoneme Mapping function would add a consonant, respectively nasal or belonging to the "r" family. Affricates and diphthongs are compared both with single L1 phonemes and with phoneme pairs. For example, all English diphthongs should be mapped into vowel pairs in Italian but can be mapped into diphthongs or vowel pairs in German.

3. RESULTS AND DISCUSSION

By informal evaluation of the foreign text pronunciation by the different Loquendo voices, we may conclude that our Phoneme Mapping algorithm yields satisfying results. Best results are obtained when L1 and L2 belong to the same linguistic family, but also in the other case the mapping is mostly convincing. The worst performance is obtained when Chinese or Swedish are involved, as can be expected due to their tonal features, which are not represented in our articulatory descriptions. A further intrinsic limitation of the approach is due to the nature of the unit-selection technique. In fact, 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 and consequently may produce acoustic discontinuities. For this reason, we are currently enriching some of our voices with speech material intended for foreign (English) pronunciation.

Other aspects of our approach are still being discussed. For example, we are currently applying the Phoneme Mapping at a broad-transcription level, prior to word-boundary phonetic changes and context-dependent allophone variations, but this choice is debatable. In principle, the availability of L1 allophones could provide a mapping closer to the original L2 pronunciation. 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. 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.

4. CONCLUSION

The described approach enables a monolingual TTS voice to speak any foreign language in an intelligible way. The method is efficient, entirely language independent and in principle does not require any tuning to be applied to new languages. Work is still in progress to refine some phonetic aspects and to improve its performances on tonal languages.

REFERENCES

[1] Quazza S., et al., "ACTOR: a Multilingual Unit-Selection Speech Synthesis System", *4th ISCA Workshop on Speech Synthesis, Perthshire Scotland, Sep. 2001.*

[2] W.N. Campbell,“Talking Foreign. Concatenative Speech Synthesis and Language Barrier”, Eurospeech, 2001

[3] “Handbook of the International Phonetic Association. A Guide to the Use of the International Phonetic Alphabet”. Cambridge University Press, 1999.