Introduction to Multilingual Corpus-Based Concatenative Speech Synthesis

Filip Deprez\textsuperscript{1}, Jan Odijk\textsuperscript{1,2}, Jan De Moortel\textsuperscript{1}

\textsuperscript{1} Nuance Communications International, Belgium
\textsuperscript{2} UIL-OTS, Utrecht University, the Netherlands

\texttt{(filip.deprez,jan.odijk,jan.demoortel)@nuance.com}

Abstract

This tutorial paper addresses foreign-language support in corpus-based concatenative text-to-speech systems. We give an overview of application domains where strictly monolingual speech synthesis is not sufficient and where multilingual text-to-speech is required or highly desirable. We describe two approaches to multilingual corpus-based speech synthesis: phoneme mapping on the one hand, and the creation of multilingual speech databases on the other. We list the strengths and weaknesses of both approaches.

Index Terms: speech synthesis, multilingualism, phoneme mapping, multilingual speech database

1. Introduction

A number of applications that incorporate text-to-speech technology require foreign-language support. To a certain extent, foreign-language support can be obtained by integrating several monolingual text-to-speech systems in the application, while having the application switch to the appropriate monolingual system depending on the synthesis needs. However, the description of such juxtaposed monolingual systems – where a language switch often goes along with the switch to a different voice – is not the object of this paper. Instead, we want to concentrate on multilingual synthesis based on one single voice, as required in each of the applications described further on.

There are many aspects to multilingual speech synthesis. A taxonomy of text-to-speech systems with foreign-language support has been proposed in [1]. Methods for language identification have been described, for example, in [2, 3]. [4, 5] focus on text analysis, while [6, 7] concentrate on multilingual prosody modeling. Multilingual text-to-speech systems have been described, for example, in [8-12].

The paper is organized as follows: in section 2 we describe a number of application domains where text-to-speech systems face the multilingual challenge. Section 3 focuses on phoneme mapping as a strategy to obtain multilingual synthesis. Section 4 concentrates on the development of multilingual speech databases. In section 5 we formulate a conclusion on both approaches.

2. Application domains

An application for reading running text (such as newspaper articles) is faced with the fact that many languages borrow words from other languages (Anglicisms in French: “weekend”, “parking”, “western”, “mail”, “shopping”; anglicisms in German: “update”, “upgrade”, “crash”, “downloaden”; German loan words in English: “Kindergarten”, “Leitmotiv”; etc). On top of that, native text can be interspersed with regular words from other languages e.g. “Ce cinéaste britannique est également l’auteur du script de The Queen”.

A directory assistance application dealing with residential entries encounters many proper names (person names, street names, etc) from foreign origin. This is especially the case in countries with more than one official language.

A directory assistance application for business entries is faced with the fact that business names often contain (English) loan words, or that they are composed of words originating from different languages (e.g. “ABX Logistics International – Deutschland”).

An application for reading SMS messages has to cope with text written in internet slang that often incorporates loan words. Spanglish, Denglish, Franglais, ... are terms to refer to the language blending that is typically used inside SMS messages, e.g. “Wir sind stolz und happy”, “Wir denken an euch in the land down under”, etc.

A navigation or traffic information system reading street and city names faces the multilingual challenge especially in foreign countries, where arguments in a foreign language are filled out in the open slots of carrier sentences in the native language. E.g. “Biegen Sie rechts ab in <Baker Street>”.

A music or video entertainment system reading aloud song and film titles as well as album and artist names inevitably has to deal with the multilingualism that is typical for the domain, e.g. “The English Patient, with Juliette Binoche in the role of Hana”.

3. A fast way to multilingualism: phoneme mapping

3.1. Architecture

Figure 1 illustrates a possible architecture for a corpus-based text-to-speech system taking its multilingual capabilities from phoneme mapping. A sentence containing text in several languages (e.g. “You have been listening to <Petit bonheur> by <Salvatore Adamo>”) is fed to the module that determines the language origin of the different parts in the input sentence. Subsequently these are routed to the appropriate grapheme-to-phoneme (G2P) converters. The phoneme mapping module converts the foreign transcriptions into native transcriptions and sends these to a unit selector / synthesizer that has a monolingual speech database at its disposal.

3.2. Development path

Assuming the availability of a language identifier and of grapheme-to-phoneme converters for the foreign languages to support, phoneme mapping is a fast and easy way to obtain multilingual synthesis: it is sufficient to make an inventory of
the phonemes in the source and target languages and to define the appropriate mappings.

![Diagram](image_url)

**Figure 1. Architecture for a text-to-speech system using phoneme mapping to support two foreign languages**

### 3.3. Strengths and weaknesses

Although the phoneme mapping approach certainly offers some potential, it also presents a number of challenges that are not easy, if not impossible, to overcome.

#### 3.3.1. Development cost and system size

The phoneme mapping approach is attractive from a system size and development cost point of view.

#### 3.3.2. Lacking phonemes

When looking at the phoneme inventories of Western-European languages, one can easily see that a fair amount of phonemes are shared over the different languages: /p/, /t/, /k/, /b/, /d/, /g/, etc. Although the acoustic realization may vary over the different languages, there is at least some degree of interchangeability, which obviously is a prerequisite for the success of the phoneme mapping approach.

The phoneme mapping approach becomes more difficult when facing phonemes of the foreign language for which there is no direct counterpart in the native language. To a certain extent, this problem can be solved by defining approximations coming from a different phoneme or phoneme sequence. In German for instance, one may approximate the French nasal vowel [ɲ] by means of the sequence [n ɹ].

For certain phonemes, however, there may be no counterpart at all. This is the case for instance for the English/Dutch/German sound [h] that is lacking in French and that can’t be successfully approximated by any other French phoneme. Apart from dropping the phoneme, which may be a valid approach, there is no real mapping alternative.

The severity of the problem of lacking phonemes can be well illustrated by contrasting the French and Spanish vowel inventories. In French, there are seventeen vowels, whereas in Spanish there are only five. A phoneme mapping approach that tries to cover French with Spanish phonemes needs to map multiple French vowels to one single Spanish vowel, and this can obviously create confusion and intelligibility problems at synthesis time. Assuming that both [u] and [y] are mapped to the Spanish [u] – and there probably is no real alternative to this mapping – the French minimal pairs “sous” and “su”, “tou” and “tu”, “fou” and “fût”, “boue” and “bu” all end up with the same transcription.

#### 3.3.3. Lacking phoneme clusters

It is well-known that the strength and the quality of a corpus-based concatenative synthesis system go hand in hand with the richness of the speech database. Not only should it contain a sufficient number of instances of the basic speech units (be they demiphones, diphones, phonemes, etc.), it should also contain a sufficient number of longer chunks, beyond the level of the basic speech unit, that are typical for the target language or application. These longer chunks can be phoneme clusters, syllables, words, phrases, up to complete sentences.

A speech database that is sufficiently rich, both in terms of basic speech units and in terms of clusters of basic speech units, allows obtaining synthesized speech characterized by a low splicing rate (or a high average segment length) and – since there is a positive correlation between the average length of the selected speech segments and the quality of the synthesized speech – a superior quality.

An important limitation that is intrinsic to the phoneme mapping approach is linked with the nature of the speech database: even if a given monolingual speech database contained all phonemes of a given foreign language, it is very unlikely that it would also sufficiently cover the longer chunks that are typical of the foreign language (e.g. syllables, words, etc.). As a consequence, the foreign language synthesis is very likely to be characterized by a higher splicing rate, a higher number of non-smooth concatenations and thus a lower quality than the synthesis of the native language.

### 4. The hard way to multilingualism: multilingual speech databases

Some of the above-mentioned limitations to multilingual synthesis can be overcome by creating a text-to-speech system with a multilingual speech database. The development path for such a speech database is hard and time-consuming, but the resulting speech quality is well beyond the quality of what can be obtained by means of phoneme mapping.

#### 4.1. Architecture

Figure 2 illustrates a possible architecture for a corpus-based text-to-speech system taking its multilingual capabilities from a multilingual speech database. In such a system, the unit selector is capable of taking multilingual phonetic
transcriptions as input, and of selecting appropriate speech segments from the multilingual speech database. To that effect, the phonemes of the input transcription as well as the phonemes in the speech database are enriched with a tag describing the language of origin.

Figure 2. Architecture for a text-to-speech system using a multilingual speech database to support two foreign languages

The speech database is populated with speech segments from all languages supported by the system. In a multilingual system with a primary language, the size of the subspeechbase for the native language will tend to be bigger than the size of the subspeechbases of the foreign languages, although the size of the foreign speechbases cannot go below a critical minimum. In a multilingual system without primary language, the size of the subspeechbases is more or less equivalent for all supported languages.

4.2. Development path

4.2.1. Voice talent selection

A first and very important step in the creation of a multilingual speech database is the selection of a voice talent that demonstrates good knowledge of the languages that need to be supported by the multilingual system. The evaluation of candidate voice talents can be done by means of a test recording session where the recording script is populated with a variety of stimuli for all languages. The test recording script will typically contain phonetically rich sentences (varying in length and reading difficulty), affirmative and interrogative sentences and prompts that are typical for the target applications (e.g. welcome messages for directory assistance systems, traffic information messages, navigation messages, email addresses, URLs, proper names, etc).

Tests using the resulting recordings, as well as the recording experience itself, allow answering the following key questions:

- Does the voice radiate the desired personality (e.g. friendly, happy, positive, young, fresh, inspiring confidence, etc)?
- Are voice timbre and reading style in line with the expectations?
- Is the pronunciation accurate for all languages? How strong is the accent when speaking the foreign languages? Is the voice accepted by the inhabitants of the different linguistic regions where the resulting TTS system is to be deployed?
- Are the recordings, across the different languages, characterized by consistency at all levels (timbre, volume, rhythm, intonation)?
- Does the voice talent have the ability to record during longer periods, without losing recording consistency?
- Does the voice talent demonstrate a collaborative attitude at recording time (e.g. does he/she accept to be corrected, is he/she willing to do retakes, etc).

4.2.2. Creation of recording scripts

Once a voice talent is selected, the recording scripts can be created. For each language that needs to be supported, the script typically contains several thousands of stimuli, and consists of a mix of phonetically balanced sentences and domain-specific material. It is good practice to bring the complexity of the foreign language sentences to a level that matches the language skills of the voice talent. This can be done by reducing the length of the sentences, by avoiding the presence of very infrequent and difficult words, or by limiting the number of phonemes that don’t have a counterpart in the native language.

4.2.3. Recordings

The recording phase is a tedious period, both for voice talent and voice coach. The recording of the native script generally goes smoother and faster than the recordings of the foreign language scripts, since these are characterized by a higher amount of hesitations and retakes. It is advised to have a native coach for the foreign language recordings: not only can this coach monitor the quality of the foreign language speech, he or she can also be helpful by saying the stimulus before it is spoken by the voice talent.

4.2.4. Creation of the speech database

Similar as for monolingual systems, the recorded data is enriched with aligned validated phonetic transcriptions and with phoneme boundary markers in view of compilation into a speech database. On top of this, a language tag indicating the language origin of each phoneme is also stored in the speech database.
4.2.5. Tuning and evaluation

During the tuning and evaluation phase one tries to find the optimal settings for the unit selector. By assigning different penalties to the non-respect of the language origin feature, one can either slightly encourage or strongly discourage language blending in synthesis. The optimal amount of language blending may be function, among others, of the size restrictions of the target system.

4.3. Strengths and weaknesses

4.3.1. Voice Talent limitations

If it is not an easy task to find a good voice talent for a monolingual system, it is a fortiori challenging to locate a suitable voice talent for a multilingual system. In any case, there is a practical limit to the number of languages that can be supported by the same voice. A very optimistic upper limit to the number of languages spoken by professional voice talents is probably somewhere around six. In practice it may already be very hard to find a professional voice talent speaking three or four languages to a sufficient degree. If additional languages need to be supported, one needs to fall back on the phoneme mapping approach for these.

4.3.2. Development cost and system size

The development of a multilingual speech database is a time-consuming, tedious and costly activity. For small-footprint systems, the size of the multilingual speech database may be an issue.

4.3.3. Accented or unaccented speech

With a multilingual speech database the look-and-feel of multilingual synthesis is similar to the look-and-feel of the multilingual natural speech of the voice talent. The native or foreign character of the speech one observes at recording time is reproduced at synthesis time. For certain applications it may be an advantage to produce unaccented foreign speech, for other applications the audience may profit from accented foreign language speech. Whatever the application’s requirement is, it needs to be taken into account at the time of the voice talent selection.

4.3.4. Quality

As with monolingual corpus-based text-to-speech systems it is possible to obtain very high synthesis quality with a system equipped with a multilingual speech database. This remark holds especially for multilingual systems developed for a specific domain e.g. directory assistance.

To illustrate the quality level that can be obtained we refer to a quadrilingual corpus-based text-to-speech system developed by Nuance (formerly known as ScanSoft). The system is used in a directory assistance application, and speaks English as well as the three official languages of the country where it is deployed. During the evaluation phase, a large-scale MOS test was set up. The test was executed by test subjects from the three linguistic regions, and contained both natural and synthetic speech. The test results showed that the multilingual capabilities of the voice talent and of the text-to-speech system were up to the level of satisfying the test subjects of all three linguistic regions. Moreover, the test showed that the average quality of the synthetic speech was only a few percentages below the average quality of natural speech.

5. Conclusion

In this paper we have discussed two approaches for multilingualism in corpus-based speech synthesis. The first approach, phoneme mapping, is attractive because of its relatively simple development path, but it is not always convincing with respect to resulting synthesis quality. While the second approach, the development of a multilingual speech database, can reach a high quality, it is limited in the number of foreign languages that can be supported and is characterized by a time-consuming, tedious and costly development path. In order to overcome the disadvantages of both approaches, different techniques – moving away from purely corpus-based concatenative synthesis (cf. for example [13]) – are being developed.

6. References