FROM MULTILINGUAL TO POLYGLOT SPEECH SYNTHESIS

Christof Traber[1], Karl Huber[2], Karim Nedir[2], Beat Pfister[1], Eric Keller[3], Brigitte Zellner[3]

[1]TIK, ETH-Zentrum, 8092 Zürich, Switzerland, {traber, pfister}@tik.ee.ethz.ch, http://www.tik.ee.ethz.ch
[2]Swisscom AG, CIT-CT-SPI, Postfach, 3050 Bern, Switzerland, {karl.huber, karim.nedir}@swisscom.com
[3]LAIP, University of Lausanne, 1015 Lausanne, Switzerland, {eric.keller, brigitte.zellner}@imm.unil.ch

ABSTRACT

This paper proposes a distinction between existing multilingual synthesis systems and *mixed-lingual* or *polyglot* synthesis systems. The latter should be capable of synthesising with the same voice utterances which contain foreign language words or word groups. As a first step towards polyglot synthetic speech, the design and realisation of a 4-lingual single-speaker diphone inventory is detailed. The first results show that mixed-lingual sentences can be synthesised using this inventory. Further work will focus on multilingual text analysis and prosodic modelling in order to create a complete polyglot TTS system.

Keywords: speech synthesis, polyglot text-to-speech, multilingual diphone inventory

1. INTRODUCTION

Many speech synthesis systems are reasonably intelligible today. But still, they might never reach the intelligibility and naturalness of a human speaker. On the other hand speech synthesis can make up for this deficiency with some other good features. In our view, the most outstanding feature of speech synthesis could be the ability to speak in many tongues. People who master 4 and more languages equally well are very rare. Speech synthesis might, of course, apply different voices for different languages. But this is not adequate in most cases. In fact when we look at newspapers we can find many sentences with embedded foreign language expressions or proper names. It would not be appropriate to have these sentences pronounced with different voices. Polyglot people are not only able to switch between languages but they can do so in the midst of a sentence and they can recognise when a language change is required. Still further they recognise (depending on external factors) what degree of language assimilation has to be adopted. They may decide to use complete assimilation to the primary language or assimilate only prosody and use the appropriate foreign phonetic inventory. Or they may decide to use both phonetics and prosody of the embedded foreign language expression (e. g., for quotations).

We can classify *speech synthesis* systems with regard to multilinguality in the following way:

1. **Pure monolingual speech synthesis:** Any foreign words must be adapted to the monolingual phonetic alphabet.
2. **Simple multilingual speech synthesis:** For every language, distinct algorithms and distinct databases are used. Switching language means switching the synthesis program or process. This can only be done between sentences. Language switching is usually accompanied by voice switching.
3. **Polyglot speech synthesis** (mixed-lingual speech synthesis with a preselected primary language): The main language of a document is identified as the primary language of the synthesiser (either automatically or preselected). The synthesiser uses this as the default language until a foreign language word or expression is encountered, which will be pronounced appropriately. Some assimilation may take place. Intonation will be more or less assimilated to the primary language. Foreign language words will be recognised by lexical means, phonotactic/compositional analysis and by using subgrammars for recognising foreign language constructs.
4. **Polyglot speech synthesis with language detection** (no preselected primary language): The text analysis module is able to detect the current language by itself using multilingual text analysis (maybe supported by statistical means). It uses appropriate phonetic and intonational models to generate utterances.

Multilingual systems of type 2 seem of limited use for applications that we have in mind. For many applications – especially in a multilingual country like Switzerland – it would be desirable to have a mixed-language speech synthesis of type 3 (termed **Polyglot Speech Synthesis**). As an example, consider an automatic reverse telephone directory service that has to synthesise a basically German sentence containing French and Italian proper names*: "Der Teilnehmer ist François Lejeune, via Roggiana 16, 6945 Origlio."*. In this mixed-lingual sentence *"Der Teilnehmer ist"*, "*16*" and "*6945*" would be pronounced in German, "*François Lejeune*" in French, and "*via Roggiana*" and "*Origlio*" in Italian.

Type 4 multilingual synthesis would be used in applications with no a-priori knowledge of the text document language or for documents with heavily mixed languages. Other example applications for *polyglot speech synthesis* are e-mail reader, audio web browser, automatic cinema program announcement.[1]

As a first step towards polyglot speech synthesis, a joint project of ETH Zurich, University of Lausanne, and Swisscom was finished last year. It consisted of the creation of a diphone inventory for German, French, Italian, and English. This year a second stage has been initiated with the collection of a prosodic database for German and French.

This paper focuses primarily on the first development phase of the polyglot speech synthesis system, namely the automatic extraction of the multilingual diphone inventory and related issues.

2. ISSUES IN POLYGLOT SPEECH SYNTHESIS

The main problems to be addressed within such a project are the following: unified treatment of differing languages within a single environment (e. g., handling of different phonetic inventories), text analysis and language recognition for mixed-language input, prosodic modelling of mixed languages, handling of a large polyglot diphone database. We will point out briefly some problems regarding databases and multilingual phonetic inventory.

2.1 Shared and loadable databases

As mentioned above, it was a design goal to have an integrated system for all languages, i. e., to have shared algorithms and either shared or loadable databases for individual languages. For example, there will be a single lexicon (full-form, morphemic, submorphemic) with language tags such that secondary language expressions can be recognised by means of lexical analysis. The lexical language tags will trigger the syntax analysis to utilise local subgrammars to detect local foreign language constructs. The sentence grammar for the primary language will be preloaded only once. On the other hand, it may be necessary to dynamically load syntax-based transformation rules to handle, e. g., French mandatory and optional *liaison*.

2.2 Multilingual phonetic inventory

When integrating different languages into a single system we would like to have a standard representation for any kind of information. The first problem arises with phonetic alphabets: Although there is a common phonetic alphabet for western languages (*IPA*), this does not imply that the same phonetic symbol used in different languages designates the same phonetic sound. E. g., the [e]-sound in the English word *"bed"* is much more *lax* than [e] in the German word *"Methan"*. Secondly, the most common pronunciation dictionaries diverge concerning their phonemic versus phonetic transcription standards. While in French phonemic transcription prevails over phonetic transcription (no distinction of long versus short phones) German dictionaries tend to give a mixed phonemic/phonetic transcription (distinction of position-dependent *allophones*). In Italian an allophonic transcription with short and long consonants is predominant (*mamma - rame*). There is a design decision to be made in a multilingual system: Do we retain the common transcription standard as used in these dictionaries or do we opt for a phonetic alphabet that distinguishes the sounds of all languages? We decided to opt for the first way in order to ease the lexicon acquisition process.

3. POLYGLOT DIPHONE INVENTORY

3.1 Compilation of carrier word lists

In a first approach we decided to use diphones for concatenative speech synthesis. At a later time we could still move towards more general units like *CVC* and *consonant clusters*. The natural language carrier utterances should be selected in a way that there could always be a fallback on diphonetic elements. This way diphone synthesis would be the minimal quality-standard for speech synthesis. Carrier word selection was based on lists of manually and automatically transcribed words (e. g., 120'000 for German). To generate phonetic transcriptions for Italian, transcription rules were implemented by means of finite-state transducers. In a first approach, a *greedy algorithm* was used to have as many diphone elements as possible with a minimum number of carrier words. It turned out that words selected this way tended to be uncommon and awkward. We then decided not to insist on a minimum number of carrier words. In a second stage we tried to complement missing diphones by construction of compounds and artificial words. In a third stage all carrier words were checked: Diphones in accented position avoiding the initial and final syllable were preferred. In total, 5300 German carrier words with 2500 diphones and 3300 Italian words with 1400 diphones were collected at TIK, 1700 French words with 1600 diphones and 4400 English words representing 2200 diphones at LAIP.

3.2 Speaker selection and recordings

Concatenative speech synthesis requires speech units to be extracted from natural speech. In our case this meant that a speaker had to be found with excellent language skills and consistent articulation in four languages. Moreover, the speaker's voice should be well suited for the prosody manipulations required in the synthesis process, and the synthetic voice should sound as pleasant

[1] See [1] for an earlier research project on polyglot speech synthesis and [2] for other aspects of foreign language speech synthesis.

as possible. Twenty-five (mostly female) candidate speakers were found and evaluated at LAIP and at TIK, the lack of foreign accent in the four languages being the most important criterion.

Six female speakers were selected for further test recordings under studio conditions. In addition to general texts, the recording material included words carrying all the diphones needed to synthesise five test sentences. These diphones were extracted manually and the test sentences were synthesised using natural prosody. The evaluation of these synthetic sentences together with the read-aloud general texts led to the selection of our final speaker. The recordings of the entire diphone carrier lists in all languages were done in several sessions during two weeks, and the speaker read the word lists in a highly monotonous style. The total raw recording time was 24 hours.

3.3 Diphone extraction

The diphone extraction was done in four essential steps, which will be explained in this section:

1. semi-automatic annotation of carrier words in the raw recordings
2. segmentation of the carrier words into phones according to the given transcription
3. extraction of the required diphones from the corresponding double phones
4. selection of the best variant of each diphone if several versions existed

3.3.1 Annotation of carrier word recordings

The annotation of the raw recordings consisted of finding the carrier words in the sound files and labelling them with the item number of the corresponding textual representation of the carrier word. A simple semi-automatic tool was created for doing this task efficiently. This program worked by repeatedly finding the next relevant stretch of audio data and playing this audio item while simultaneously displaying the next textual item. A set of correction commands allowed the user to skip over spurious sound items (such as coughs or commenting remarks), to enlarge the sound portion (e. g., if the end point was wrongly detected as lying in the occlusion phase before a final plosive) or to find the appropriate textual item forward and backward in the carrier word list (used especially to annotate corrections or repetitions of carrier word readings). All in all, the annotation process did not take much more time than the real studio recording time.

3.3.2 Phone segmentation and diphone extraction

The *French diphones* were extracted at LAIP. This extraction process started by a high-quality manual segmentation of the carrier word into phones at positions where a diphone was to be extracted. Different people contributed to this segmentation, and it was rigidly controlled that the same segmentation criteria were applied by all of them and throughout the segmentation task. In a second step, diphone boundaries for stationary sounds were automatically placed in the middle of the manually segmented phones. For diphones starting or ending in a plosive, the beginning of the burst was manually labelled as the diphone boundary.

For German, English, and Italian, a *fully automatic diphone extraction* was developed and applied at TIK, which consisted of an HMM-based segmentation of the carrier words into phones and a subsequent determination of diphone boundaries and selection of the best variant of each diphone.

The *phone segmentation* was obtained by applying an embedded HMM-training using all the carrier material for one language. No gross errors were found in the final segmentation in a test sample, i. e., no errors in which a phone boundary was misplaced beyond the middle of the next or previous phone, which was a major prerequisite for a successful determination of diphone boundaries.

The automatic determination of diphone boundaries in stationary sounds and the selection of the best of several variants of a diphones was based on cepstral distance measures between frames of carrier word phones and a *phone centroid*, i. e., an average typical cepstrum of the phone in question. For each stationary sound of each language such a centroid was computed. This process is exemplified here for the sound [a:]: An initial centroid was created by taking the average of all cepstral vectors (20 coefficients) of all frames of all [a:] segments which belonged to a diphone to be extracted. Then, the centroid was refined iteratively as follows: For each [a:], the frame with minimal cepstral distance to the current centroid was determined, and a new centroid was computed by taking the average of all these minimum-distance cepstral vectors. This process was repeated until convergence occurred, i. e., until the minimum-distance frame positions did not change any longer.

The *determination of diphone boundaries* was then carried out in the following manner:

For stationary phones, the frame with minimal cepstral distance to the corresponding centroid was taken as the diphone cutting point. At least, this was the original intention, with the aim to minimise the spectral distortions between all diphones starting and ending in a specific sound. We found, however, that this distance measure alone was not reliable enough, and we then simply limited the range where the minimum distance (and hence, the cutting point) was to be found to the interval from 40% to 60% of the phone.

For unvoiced plosive sounds, the cutting point was defined as the frame immediately before the burst. This point could be found very reliably by minimising the empirically found function

$$d_i = -[\log(rms_{i+1})-\log(rms_i)] / \mathrm{sqrt}(rms_i)$$

where rms_k represents the RMS energy of frame k (applied on a pre-emphasised signal).

Cutting points for voiced plosives were treated similarly to unvoiced plosives. However, the simple function

$$d_i = -[\log(rms_{i+1})-\log(rms_i)]$$

was minimised to find the cutting point.

Finally, for all diphones with more than one realisation, the *best variant* was selected based again on the cepstral distance between the diphone boundary frames and the corresponding centroids (in the case of stationary sounds). For example, from all versions of the diphone [a:n], the one with minimal summed-up cepstral distance between the left and right boundary frame and the centroids for [a:] and [n] was chosen as the best diphone. For diphones including plosive sounds, only the boundary distance of the non-plosive part was used as selection criterion.

4. RESULTS AND EXAMPLES

The automatically extracted diphones for German, English, and Italian together with the French diphones from LAIP form a first version of a complete quadrilingual diphone inventory, with which it is possible to synthesise monolingual as well as mixed-lingual sentences. Currently, it is only possible to synthesise test sentences off-line based on a manually established list of diphones and prosody values taken from natural utterances. In order to evaluate the automatic diphone extraction, a German test sentence was synthesised using automatically and manually extracted diphones from the same carrier words. Unfortunately, both versions exhibited a large amount of discontinuities at concatenation points, and the quality of the synthesis did not seem satisfactory.

In order to overcome the concatenation discontinuities to some extent, a simple and preliminary *time-domain smoothing method* was applied, which could easily be integrated in the applied TD-PSOLA synthesis procedure: At each boundary between two diphones, the last part (30 %) of the first diphone was gradually adapted to the beginning of the next diphone by applying a weighted averaging of the windowed pitch periods in the last part of the first diphone and the first windowed pitch period of the second diphone.

A longer example of mixed-lingual synthesis (a cinema title listing such as it could appear in a Zurich newspaper) is available in the CD-ROM proceedings. Further commented synthesis examples can be found at *http://www.tik.ee.ethz.ch/~spr/*.

5. FUTURE WORK

Priority was given to making a first 4-lingual single-speaker diphone inventory available. Of course, possible improvements of this concatenative polyglot synthesis will be investigated, including

- improvement of the automatic boundary placement of phones and diphones
- use of non-diphone speech units (e. g., CVC clusters to avoid concatenation points within vowels)
- dynamic unit selection, as already implemented in several concatenative TTS systems
- better spectral smoothing using frequency-domain synthesis methods.

As already mentioned before, the creation of a multilingual single-speaker diphone inventory is only the first of several steps that will lead to a polyglot speech synthesis system.

The second phase, which is currently in progress, concerns the development of prosodic models for each of our 4 languages and the integration of the quadrilingual diphone inventory in the SVOX TTS system [4]. The resulting prosodic database will be used to train statistical models for the predictions of F0 contours [4] and phone durations [3].

6. CONCLUSION

In this paper, we have first shortly described what we mean by a *polyglot* speech synthesis system as opposed to a *multilingual* system.

Our first step towards a polyglot synthesis system – namely the production of a 4-lingual, single-speaker speech unit inventory – has then been presented. In an effort to lower the cost of producing new voices, an automatic diphone segmentation procedure has been implemented. Preliminary informal listening tests suggest that a manual diphone segmentation of our diphone carriers would only lead to a concatenative speech of comparable quality, but at a much higher cost.

Together with the improvement of diphone selection and concatenation, much more work will be needed in the areas of text analysis and prosody modelling until our system can be used in a polyglot context.

7. REFERENCES

[1] Boves, L. (1991), Considerations in the design of a multi-lingual text-to-speech system, *Journal of Phonetics*, 19, pp. 25–36.

[2] Campbell, N. (1998), Foreign-language speech synthesis. *Proceedings 3rd ESCA/COCOSDA Internat. Workshop on Speech Synthesis*, Jenolan Caves, Australia.

[3] Riedi, M. (1998), Controlling segmental duration in speech synthesis systems. *Diss. ETH Nr. 12487*, Swiss Federal Institute of Technology, Zurich.

[4] Traber, Ch. (1995), SVOX: The Implementation of a Text-to-Speech System for German. *Diss. ETH Nr. 11064*. Swiss Federal Institute of Technology, Zurich.