HIGH-QUALITY SPEECH SYNTHESIS FOR PHONETIC SPEECH SEGMENTATION

F. Malfrère and T. Dutoit

Circuits Theory and Signal Processing Lab

Faculté Polytechnique de Mons

31, Boulevard Dolez, 7000 - Mons (BELGIUM)

Tel. +32.65.37.41.33, FAX: +32.65.37.41.29, Email: {malfrere, dutoit}@tcts.fpms.ac.be

ABSTRACT

This paper presents an original technique for solving the phonetic segmentation problem. It is based on the use of a speech synthesizer for the alignment of a text on its corresponding speech signal. A high-quality digital speech synthesizer is used to create a synthetic reference speech pattern used in the alignment process. This approach has the great advantage on other approaches that no training stage (hence no labeled database) is needed. The system has been mainly evaluated on French read utterances. Other evaluations have been made on other languages like English, German, Romanian and Spanish. Following these experiments, the system seems to be a powerful tool for the automatic constitution of large phonetically and prosodically labeled speech databases. The availability of such corpora will be a key point for the development of improved speech synthesis and recognition systems.

1. INTRODUCTION

Concatenative speech synthesis techniques currently achieve a high segmental quality. Among these techniques, we can highlight some recent diphones concatenation-based synthesis techniques (see for instance the publicly available MBROLA [1] speech synthesizer at URL address `http://tcts.fpms.ac.be/synthesis`) or the very promising technique of non uniform units concatenation [2],[3]. With the increase of the memory and computation power of computers, it is likely that speech synthesizers with higher segmental quality can still be developed in the years coming.

To obtain natural-sounding speech, speech synthesizers must be provided with a correct phonetic string and a natural prosody. In text-to-speech systems, the phonetic string can be derived from text either by lexicon look-up or by rules designed by an expert or obtained by training a phonetizer on a large phonetized corpus (see [4], Chap. 5 for a review).

For prosody generation, finding a natural rhythm and a correct intonation is not a trivial problem. Automatic prosody generators cannot yet deliver the expected high-quality prosody. This explains the limited naturalness of commercially available text-to-speech systems and the rapidly expanding research community on prosody generation problems [5].

Recent research in automatic prosody generation seems highlight a potential interest for machine learning techniques such as linear regression [6], classification and regression trees (CARTs) [7], neural networks (like in [8]) or statistical approaches [9]. All these techniques are based on the automatic analysis of large prosodically and phonetically labeled corpora.

The phonetic speech segmentation system presented in this paper provides an important shortcut for the creation of such databases. It performs the alignment of a text on the associated speech signal. The output of the system is a rich phonetic transcription composed of phoneme labels (derived from the text) associated with phone duration and pitch measured on the speech signal. No tedious manual segmentation is needed to label the speech corpora.

The paper is divided as follows. In Section 2, we describe the originality of our approach and compare it with other existing techniques. Our text-to-speech alignment system based on a speech synthesizer for the phonetic segmentation is developed in Section 3. In Section 4, some results are given and commented. Applications of such an alignment system are described in Section 5. The article ends with some prospects on the development of the system.

2. TEXT-TO-SPEECH ALIGNMENT APPROACHES

Speech segmentation into phonetic units is one of the major problems encountered in phoneme-based speech recognition systems. In the context of text-to-speech alignment, where the sequence of phonemes is known, the segmentation task is much more constrained, hence much easier. Several authors have already described the use of hidden Markov models (HMMs) for forced alignment of speech on text [10],[11]. These approaches, typically based on phoneme models, present advantages and drawbacks. They acquire some spectral knowledge of phonemes by analyzing large databases and are intrinsically speaker independent but the information about phoneme transition is poor. A means of tackling

this problem would be use context-dependent phoneme models. In this case, however, since nothing ensures that the phoneme boundaries estimated by the HMM training algorithm will correspond to real phonetic boundaries (since HMMs are then trained on data with contexts and the boundaries obtained are simply based on the maximization of some likelihood). Thus, such an approach would require supervised training, i.e. very large labeled speech corpora (several hours) that are not always available.

To avoid the major drawbacks of the need of a large labeled database, the approach of text-to-speech alignment described here (see also [12], [13]) is radically different. A high quality concatenative speech synthesizer is used to produce a synthetic reference signal from the phonetic transcription derived from the text. The speech signal is then temporally aligned on this reference in which the phonetic segmentation is known. The alignment process is reduced to a simple dynamic time warping algorithm [14]. In comparison with HMMs, this approach apparently looses the speaker independence since only one voice is used as reference. In practice, however, we found when testing the algorithm on several different voices that the much better segmental representation used here counterbalances the dependence on a single reference voice. As shown in Section 4, two different voices (a man and a woman voice) must be used to obtain a correct alignment in any case.

3. SPEECH SYNTHESIS AND SEGMENTATION

Figure 1 shows a block diagram of a text-to-speech alignment system based on a speech synthesizer.

The phonetic transcription is automatically derived from the input text by using an automatic text-to-speech phonetization system. For the tests reported here, we used the LIPSS system (cf. [15]) which incorporates a morpho-syntactic analyzer to provide accurate phonetic transcription of sentences.

The MBROLA diphone-based speech synthesizer is then used to create the reference speech signal (sampling frequency of 16 kHz). It should be noted that in order for the alignment to be meaningful, the synthesizer used should be implemented so as to synthesize exactly the amount of speech expected. This is precisely the case for the MBROLA synthesizer, which carefully avoids timing error accumulation.

Although natural prosodic information is needed to deliver natural sounding speech, a very rough prosody can be used to obtain the reference signal since only its segmental features will be used in the alignment process.

Phoneme duration and intonation contours are chosen so as to facilitate the alignment process.

The phoneme duration is correlated with the local continuity constraint used during the dynamic time warping segmentation stage. A constant duration of 100 ms has been chosen. This duration allows minimum and maximum phone duration of about 33 ms and 500 ms respectively.

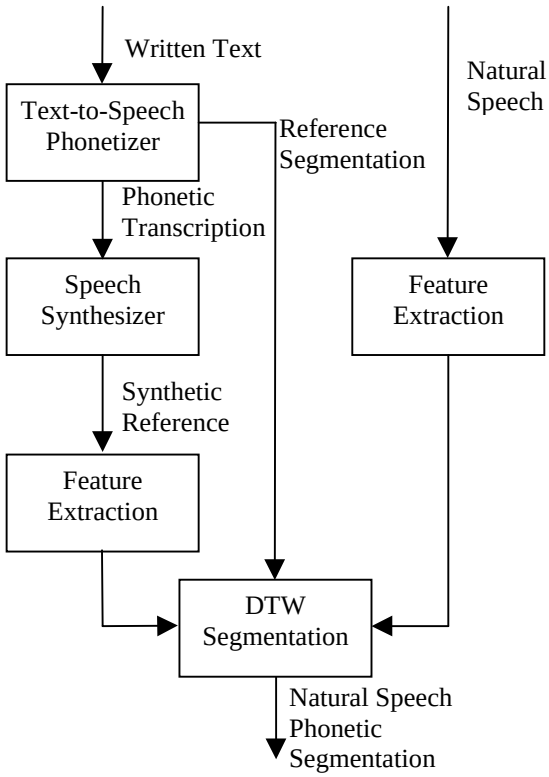

Figure 1 – Phonetic Segmentation Process

The F_0 contour used to generate the reference has been chosen as simple as possible (constant F_0 value) since no assumption can be made on the contour actually produced by the speaker. Assuming that the features extracted to compare the reference and the natural speech signals are independent of the F_0 contour (which is only approximately true (see [4], Chap. 7) but at least true enough for TTS alignment), this has no effect on the accuracy of the segmentation.

The next step corresponds to the extraction of relevant features from both the reference and the natural speech signals. Four classical sets of parameters have been used to characterize each speech frame. The first series of parameters defines a representation of the local speech spectral envelope : the cepstral coefficients (c_i) obtained from a linear prediction analysis of the frame. These parameters are weighted with a sinusoidal function in order to increase their robustness to noise [16]. The temporal derivative coefficients (dc_i) are computed to account of their temporal evolution. These derivative form the second set of coefficients. Finally, the energy (E) and the delta energy (DE) normalized coefficients are added to the two first sets of coefficients.

The last step of the process is the segmentation itself. It is realized with a classical dynamic time warping algorithm. The distance used to compare a frame from the synthetic reference and a frame from the natural

speech signal is a weighted combination of a cepstral distance and an energy distance (see Equation 1).

$$d(a,b) = \alpha . \sum_{i=0}^{OCep} (c_i(a) - c_i(b))^2 +$$
$$\beta . \sum_{j=0}^{OCep} (dc_j(a) - dc_j(b))^2 + \gamma . (E(a) - E(b))^2 +$$
$$\varphi . (dE(a) - dE(b))^2$$

Equation 1 – Weighted distance for the alignment

An optimization phase over the different parameters used in the distance has led to the following values:

- Frame of 30 ms with a overlap of 10 ms
- Linear predictive analysis order: 12
- Cepstral analysis order: 18
- $\alpha = 1.0$; $\beta = 1.25$; $\gamma = 0.75$; $\varphi = 1.25$

The next section shows the result of the first experiments performed with this segmentation system.

4. RESULTS

A prosody transplantation system [17] has been used to perform a preliminary perceptual test. In the test, a listener compared the original and the synthetic speech signals. The prosody transplantation tool precisely allows the user to quickly detect segmentation errors: in V/UV transition zones, especially, shifted phoneme often boundaries imply to produce a voiced phoneme with a totally wrong pitch value. When differences were found in the rhythm of the two signals, phoneme boundaries were hand corrected. Each move, independently of its amplitude, counted for one segmentation error. This evaluation has been made on 40 French read sentences pronounced by 5 male speakers. Results are given in Table 1.

Speakers	Nb. of Phones	Error Rates
SPK1	1033	7.9 %
SPK2	1033	5.6 %
SPK3	1036	8.4 %
SPK4	1046	7.2 %
SPK5	1036	8.5 %

Table 1 – Preliminary test results

A more reproducible and objective test has been made on the same data. The segmentation given by the system was compared to a manual segmentation. The segmentation errors were ranked as a function of their amplitude (cf. Table 2). WM1 and WM2 are two female voices (same sentences as the others speakers), the result obtained with the male voice of the synthesizer shows that, as expected, segmenting a female voice requires a female voice for the creation of the synthetic reference speech signal.

Speakers	<10	<20	<30	<40	<50	>50
SPK1	69.7	84.4	89.0	91.3	93.3	6.7
SPK2	66.7	80.0	84.6	88.4	91.0	9.0
SPK3	65.9	80.0	84.9	89.1	91.5	8.5
SPK4	60.3	74.3	79.8	84.2	88.3	11.7
SPK5	70.5	84.1	87.7	89.9	92.5	7.5
WM1	29.2	56.3	72.8	82.5	88.2	11.8
WM2	38.8	70.0	84.4	90.3	92.9	7.1
BASESPK	83.7	96.1	97.1	97.7	98.0	2.0

Table 2 – Error rates (%) as a function of the segmentation amplitude error in ms

A last experiment has been made for a male telephone voice (40 same sentences), which implied to low-pass filter the synthetic reference and to down-sample it to 8 kHz. The alignment parameters used for this test were the following (they have not been optimized):

- Frame of 30 ms with a overlap of 10 ms
- Linear predictive analysis order: 10
- Cepstral analysis order: 12
- $\alpha = 1.0$; $\beta = 1.25$; $\gamma = 0.75$; $\varphi = 1.25$

Speaker	<10	<20	<30	<40	<50	>50
FM1	26.4	48.6	63.7	74.5	80.2	19.8

Table 3 – Telephone male voice alignment results

Following these results, it seems that the system leads to good results for natural voices of the same quality as the synthesizer diphones database.

Finally, some preliminary results on other European languages (10 read sentences) are given at Table 4.

Language	<10	<20	<30	<40	<50	>50
English	57.9	75.5	83.4	86.5	89.2	10.8
German	71.3	76.8	79.1	80.6	81.4	18.6
Romanian	67.4	82.6	89.4	93.9	95.4	4.6
Spanish	86.0	93.1	94.6	96.0	96.6	3.4

Table 4 – English, German, Romanian and Spanish preliminary results

5. APPLICATIONS

The main application of the segmentation system is the creation of large labeled speech corpora. Such corpora are needed for speech recognition systems but also for speech synthesis based on unit selection [2], [3].

If the alignment system is associated with a pitch extractor, it can be used as a prosody transplantation tool ([10][17]). Such a transplantation tool may be used to evaluate the segmental quality of speech synthesizers when provided with a natural prosody directly copied from human reading. It can also be used as an ultra low bit rate speech coding system for special applications where the identity of the speaker is not important.

Another application is the use of the alignment system to accelerate the creation of new voices for existing speech synthesizers given a first set of synthesis units [18].

Last but not least, it will be of considerable interest for future research on prosody generation based on the analysis of very large phonetically and prosodically labeled corpora.

6. CONCLUSIONS AND PROSPECTS

Many applications can take advantage of such an automatic segmentation system. Following the results of these first experiments, it seems that the use of high quality speech synthesis in the context of automatic phonetic segmentation leads to low segmentation error rates. This is undoubtedly due to the fact that the reference signal provided to the dynamic time warping algorithm gives a good segmental representation of the speech signal to be aligned with the input text. This advantage of our approach over the use of HMM techniques seems to be sufficient to counterbalance, at least partially (i.e., except for sex dependency), the intrinsic speaker dependency of the approach.

Preliminary experiments have shown that this segmentation technique is can be successfully applied to other languages (only the speech synthesizer database changes) and leads to similar error rates. This means that the system could be probably efficiently used for the creation of multilingual phonetic/prosodic databases.

A very interesting research direction could be to substitute the deterministic TTS phonetizer used here with a probabilistic phonetizer specially designed for the problem of text-to-speech alignment. Such a phonetizer should provide a phonetic lattice with different possible pronunciations, possibly rated as a function of their likelihood. The alignment system would then be in charge of finding the correct transcription in the lattice, as well as the best alignment path. Some preliminary experiments have been made and point out that the system is able to find the exact transcription of the speech.

7. ACKNOWLEDGMENTS

The authors are especially grateful to the F.R.I.A. (Funds for Formation and Research in Industry and Agriculture) for its financial support.

REFERENCES

[1] T. Dutoit, V. Pagel, N. Pierret, F. Bataille & O. van der Vrecken, *« The MBROLA Project : Towards a Set of High Quality Speech Synthesizers Free of Use for non commercial purposes »*, Proc. ICSLP96, pp. 1393-1396, 1996

[2] A. J. Hunt & A. W. Black, *« Unit Selection in a Concatenative Speech Synthesis System using Large Speech Database »*, Proc. ICASSP96, pp. 373-376, 1996

[3] A.W. Black & N. Campbell, *« Optimizing Selection of Units from Speech Databases for concatenative Synthesis »*, Proc. Eurospeech95, pp. 581-584, 1995

[4] T. Dutoit, *« An Introduction to Text-to-Speech Synthesis »*, Kluwer Academic Publishers, 1997

[5] Y. Sagisaka, N. Campbell & H. Higuchi, *« Computing Prosody: computational models for processing spontaneous speech »*, Springer Verlag Edition, 1997

[6] A.W. Black & A.J. Hunt, *« Generating F_0 contours from ToBI labels using linear regression »*, Proc. ICSLP96, pp. 1385-1388, 1996

[7] M. D. Riley, *« Tree-Based Modelling for speech Synthesis »*, Proc. Second ESCA/IEEE Workshop on Speech Synthesis, pp. 229-232, 1994

[8] C. Traber, *« F0 Generation with a Database of Natural F0 Patterns and with a Neural Network »*, in Talking Machines : Théories, Models, and Designs, G. Bailly and C. Benoît, eds. North Holland, pp. 287-304, 1992

[9] B. Möbius, M. Pätzold & W. Hess, *« Analysis and synthesis of German F_0 contours by means of Fujisaki's model »*, Speech Communication, pp. 53-61, 1993

[10] D. Talkin & C. W. Wightman, *« The Aligner : Text to Speech alignment using Markov models and a pronunciation dictionary »*, Proc. Second ESCA/IEEE Workshop on Speech Synthesis, pp. 89-92, 1994

[11] B. van Coile, L. Van Tichelen, A. Vostermans, J. W. Wang and M. Staessen, *« PROTRAN: A Prosody Transplantation Tool for Text-to-Speech Applications »*, Proc. ICSLP94,1994

[12] F. Malfrère & T. Dutoit, *« An Alignment System for Prosodic Parameter Extraction of a French Text »*, Proc. Datalinguistisk Forenings 96, Aalborg, pp. 139-150, 1996

[13] P. Di Cristo & D. Hirst, «Un procédé d'alignement automatique de transcriptions phonétiques sans apprentissage préalable», to appear in Proc. 4ème Congrès Français d'Acoustique, 1997

[14] L. R. Rabiner & B. H. Juang, *« Fundamentals Of Speech Recognition »*, Prentice Hall Signal Processing Series, 1993

[15] T. Dutoit, *« High Quality Text-to-Speech Synthesis of the French Language »*, Ph.D. dissertation, Faculté Polytechnique de Mons, 1993

[16] B. H. Juang, L. R. Rabiner & J. G. Wilpon, *« On the use of Bandpass Liftering in Speech Recognition »*, IEEE Transactions on Acoustics, Speech and Signal Processing, pp. 947-954, 1987

[17] F. Malfrère & T. Dutoit, *« Speech Synthesis for Text-to-Speech Alignment and Prosodic Feature Extraction »*, to appear in Proc. ISCAS97, 1997

[*18*] T. Portele, K.-H. Stöber, H. Meyer & W. Hess, *« Generation of Multiple Synthesis Inventories by a Bootstrapping Procedure »*, Proc. ICSLP96, pp.2392-2395, 1996