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

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

CHOOSE THE BEST TO MODIFY THE LEAST: A NEW GENERATION CONCATENATIVE SYNTHESIS SYSTEM

Marcello Balestri, Alberto Pacchiotti, Silvia Quazza, Pier Luigi Salza and Stefano Sandri

CSELT - Centro Studi e Laboratori Telecomunicazioni S.p.A.
Via G. Reiss Romoli 274, 10148 Torino, Italy
silvia.quazza@cselt.it

ABSTRACT

The paper describes a corpus-based approach applied in the evolution of ELOQUENS®, the CSELT text-to-speech synthesis system for Italian, towards multi-voice, multi-language, high-naturalness concatenative synthesis. The acoustic modules have been redesigned, according to the idea of reducing the number of junctions and the need of prosodic modification. Appropriate phonetic coverage methods were applied in the acoustic database design. Automatic processing tools performed phone and diphone segmentation, pitch marking, prosodic feature detection. The synthesis algorithm exploits the speech material at its best, searching for the longest suitable sequences in the database, according to weighted distance measures on phonetic/prosodic parameters. Signal modification techniques are applied only if necessary, to smooth residual prosodic jumps at unit boundaries. The resulting voice is quite human-sounding.

Keyword: corpus-based concatenative synthesis

1. INTRODUCTION

Voice technology applications have created a growing demand for multi-lingual, multi-voice, multi-style speech synthesis systems and, first of all, for a natural sounding voice, close to the quality of pre-recorded speech. On the other hand, the computational difficulties limiting traditional speech synthesis have been largely overcome by present computers, and robust automatic speech analysis and labeling tools are now available. Such a background has suggested the idea underlying new-generation speech synthesis systems: reaching quality through quantity, i.e. providing the system with multiple external knowledge bases and obtaining high-quality voice timbre by concatenating units longer than diphones and available in many prosodic variants [1].
The recent developments of speech synthesis research at CSELT can be set in this framework. Demands for multi-voice multi-language human-sounding synthesis have imposed to improve on ELOQUENS®, the CSELT text-to-speech system for Italian [2]. The acoustic modules of the system have been entirely redesigned around the idea of corpus-based synthesis. The new system is functionally equivalent to ELOQUENS® and is replacing it in the applications developed at CSELT [3]. The acoustic database, in its current realization, is about 40 times as big as the ELOQUENS® diphone dictionary and computation time is higher, but the system is still quite suitable for real-time multi-channel applications. The improvement on the standard diphone voice is striking and the possibility of integration with prerecorded speech and with domain-dependent speech material offers a further advantage for its application in interactive voice services. The following paragraphs outline the features of the system and its underlying principles.

2. THE CORPUS-BASED APPROACH

Perceptual distortions due to relevant prosodic modification of the original signal and acoustic discontinuities due to frequent junctions are the major drawbacks of traditional diphone-based concatenative text-to-speech synthesis. Reducing signal modifications and number of junctions is the goal of corpus-based techniques, where static small-size diphone dictionaries are replaced with open structure large speech databases, covering widespread phonetic sequences in various prosodic contexts. The key factors in this approach are the *phonetic and prosodic coverage* of the intended domain and the run-time *selection criteria* for the acoustic units. If long acoustic units with prosodic features matching the input requirements are present in the database, the need of prosodic adjustment is reduced, or even eliminated. The principle has been applied, for example, in a specialized ELOQUENS® version intended for a telephone reverse directory service application [4], where the limited domain, names and addresses read as isolated words, could be covered with long positional units to be merely concatenated. When extending this approach to unrestricted text-to-speech synthesis, a trade-off is necessary between dictionary size (domain coverage) and signal adjustments. The following strategy was adopted for the newly developed system: a) to rely on labeled acoustic databases from which to extract the longest and best fitting units, without pre-defining their size and nature, and b) to apply prosodic adjustments only where necessary. The approach required a complete redesign of the synthesis algorithm and a redefinition of the role of prosody. The new algorithm is data-driven.

The available speech material is searched for the most suitable phonetic sequences, according to a best score procedure based on weighted distance measures between data and input context. Unit junctions are preferably forced at easy-to-concatenate boundaries. In most cases synthesis units are connected through simple waveform concatenation, whereas prosodic modifications of the original speech signal are performed only at few pitch jumps of relevant size or when a suitable prosody is not available in the database. Rather than an artificial pattern superimposed on the selected speech signal, prosody is now a crucial aspect to evaluate when comparing the input requirements and the available acoustic data. The definition of prosodic categories may be helpful in this respect, allowing to classify data according to their linguistic-prosodic role and to compare them with a functional description of the prosodic synthesis target. In order to obtain the greatest generality for the synthesis algorithm, no assumption is made on the contents of the database. Any speech material can be turned into an acoustic dictionary by means of automatic tools that label it with fine-grained phonetic and prosodic information. The minimal unit in such analysis is the *demi-phone*, defined as the signal portion delimited by a phone boundary and a diphone boundary. Such unit is small enough to give a detailed picture of prosody and refers both to the linguistic concept of phoneme and to the typical junction point for concatenative synthesis. The general and automatic treatment of acoustic databases, together with the clear separation between data and algorithms, is a condition for fast prototyping of new voices and languages and for acoustic specialization, where the general-purpose database is enhanced with application-dependent speech material.

3. CORPUS DESIGN

Although the implemented algorithm is quite general and can exploit at its best any speech material, a careful design of the texts to be recorded is necessary to take advantage of long and suited acoustic units [5]. A good phonetic and prosodic coverage of the intended domain (language/application) must be ensured. The main steps in the design of an acoustic database are the following:

- definition of the target phonetic/prosodic coverage by means of statistical analysis of a large text corpus representative of the intended domain
- extraction, by means of a greedy algorithm, of a minimal subset of well-formed texts ensuring the intended coverage

A-priori prosodic coverage is defined in terms of sentence structure: phrase type and position in the phrase. What should be noted is that, for the expected prosody to be realized by the speaker, the selected sentences should be regular in structure, not too long and easy to read. The resulting corpus to be recorded will not be free from redundancies, as it is made up of real phrases and frequent phoneme sequences may occur repeatedly. When necessary (excessive size of the acoustic database) redundancies will be pruned on the basis of their actual prosodic realization (see 4.6).

4. ACOUSTIC DATABASE PROCESSING

4.1 Speaker Selection and Audio Recordings

The selection of the reference voice(s) goes through a performance test of a lists of professional speakers: several parameters are judged, such as accent, voice pleasantness, articulation, speaking rate and intonation control, waveform morphology, robustness to speech algorithms, and so on. The corpus material, organized in texts and phonetic transcriptions, is carefully read by the selected speaker(s), collected with a high quality microphone in a recording studio, and A/D converted at 16 kHz sampling frequency into speech files.

4.2 Phone and Diphone Segmentation

An automatic phonetic segmentation module, based on context independent CDHMM phone modeling [6], aligns each speech waveform with its corresponding phone labeling sequence. The acoustic analysis, performed every 5 ms, is based on the computation of 8 mel-scaled cepstral coefficients from the outputs of a 24 bandpass filter-bank. Feature vectors, phonetic transcriptions and 45 unit HMMs are processed by a Viterbi alignment method to produce phone boundaries. The average performance is over 95% in speaker dependent mode within a tolerance of 20 ms.
A rule-based diphone segmentation module exploits three acoustic parameters (signal energy, spectral variation function and relative phone duration), two conditions (equality and belonging to a range of values) and two logical operators (AND/OR). A rule parser, connected with the data structure of the phonetic aligner, produces diphone boundaries by processing a set of about 200 acoustic/phonetic rules. At this step, the *demi-phone* sequence can be identified.

4.3 Pitch Marking and Time Alignment

A proprietary pitch period detection algorithm, based on the search for maxima of a weighted modified autocorrelation function, is applied. On voiced portions of speech, pitch estimate errors are detected and corrected by a forward/backward adaptive procedure: pitch markers are assigned in correspondence of the nearest left zero crossings of waveform peaks. Voiced phone and diphone boundaries are aligned to the nearest pitch marker positions, whereas unvoiced phones are labeled with equally spaced intervals.

4.4 Prosodic Labeling

The demi-phone sequence is labeled with some acoustic parameters of relevant prosodic importance: for each

speech portion corresponding to a demi-phone, the following parameters are computed:

- Duration in ms;
- Signal rms;
- Average F0 in Hz (voiced only);
- Average first derivative of F0 in Hz/s (voiced only).

Average F0 values of unvoiced demi-phones are linearly interpolated from the values of the nearest left and right voiced demi-phones.

4.5 Database Building

Finally, the whole material is separated into phrases, each represented in a database structure with the relevant information associated to the corresponding demi-phone constituents: speech filename, phonetic transcription, demi-phone boundaries, average F0 values, pitch markers and categorical prosodic labels, identifying phrase type and demi-phone position inside the phrase.

4.6 Database Pruning

Repetitions coming from identical positions in sentence structure may be eliminated without loss of quality for the synthetic voice. A tool has been implemented that looks for replicated units (e.g. phone sequences delimited by easy-to-join phones) and selects a typical mean representative for each set of identical units, on the basis of a prosodic distance measure taking into account duration and average F0 of each demi-phone in the unit.

5. THE SYNTHESIS ALGORITHM

5.1 Unit Selection

The core of the synthesis algorithm is the selection of the best fitting units from the acoustic database. Units are not pre-defined, rather they amount to the longest *suitable* demi-phone sequences that can be found in the database. The input to the selection algorithm is a stream of phonemes, each marked with its categorical prosodic label and its F0 and duration values [7]. The stream is scanned left-to-right, each input demi-phone is looked up in the database and its occurrences are scored according to their match with the input phonetic and prosodic context. Finally the best scored candidate is selected. The scoring procedure combines the following criteria:

(1) *phonetic context*: scores the match between the input phonetic stream and the candidate phonetic context, by means of a bell-shaped window centered on the focussed demi-phone and delimited by the first unmatched phonemes, whose degree of similarity (according to articulatory class, stress, voicing) with the corresponding input phoneme is also considered;
(2) *prosodic transition*: scores the smoothness of the transition from the previously selected demi-phone, taking into account F0 and duration values;
(3) *target prosodic type*: compares the candidate prosodic label with that of the input phoneme;
(4) *target prosodic values*: compares the F0 and duration values of the candidate with those of the corresponding input demi-phone.

Such distinct scores are combined into a weighted sum, scaled by (5) a *concatenation factor* which encourages selection of adjacent demi-phones and concatenation of units at easy-to-join boundaries. Weights can be tuned and are at present under refinement. For prosodic types having a representative in the acoustic database, a smoother and more natural sounding voice is obtained when leaving out criterion (4). This suggests that unit selection may be better driven by abstract prosodic types than by target numerical prosodic values (see 6.2).

5.2 Prosodic Adjustments at Unit Junctions

Although the unit selection process looks preferably for smooth prosodic transitions, residual discontinuities at unit junctions may still be possible, when jumps to a different source location are imposed: this may cause perceptual troubles, mainly due to large F0 differences at voiced sound boundaries. To overcome this problem, a compensation mechanism has been introduced. Given a synthesis unit, i.e. a sequence of demi-phones (i)...(j) coming from the same speech source, if the F0 jump at the junction with the preceding unit exceeds a threshold (10 Hz), a left pitch scaling factor between demi-phones (i) and (i-1) is defined as:

$$psf_l\ (i) = 100 * F0\ (i-1) / F0\ (i)$$

and a right pitch scaling factor between demi-phones (j) and (j+1) is defined, normally set to 100 except when the difference [F0(j) - F0(j+1)] has the same sign as [F0(i) - F0(i-1)] and an absolute value greater than the threshold, in which case it is set to:

$$psf_r\ (j) = 100 * F0\ (j+1) / F0\ (j)$$

A current pitch scaling factor psf_c, linearly interpolated between psf_l (i) and psf_r (j), is computed, and a proprietary time-domain pitch synchronous algorithm, CSELT-SEQUENS®, is applied to the original waveform for proportional modification of each analysis pitch period length lpa into a corresponding synthesis pitch period length lps:

$$lps = 100 * lpa / psf_c$$

In such a way, the original shape of the intonation curve of each unit is preserved, and submitted to a percentage frequency scaling which smoothes F0 jumps and realigns out-of-range units. In most cases, where F0 differences are lower than the threshold, a pure waveform concatenation is sufficient, without any signal distortion.

5.3 Optional Enforcement of Target Prosody

If the selected units do not match the target prosodic labels, so that the original prosody would be too different

from the intended one, the target prosodic values (F0 and duration) computed by the phonetic module are imposed on the original signal by means of CSELT-SEQUENS®.

6. EXPERIMENTS

6.1 Implemented Databases

A first application of the design procedure has yielded the current acoustic dictionary for the Italian language (male voice), amounting to 60,000 phonemes in 1350 declarative sentences of simple phrase structure (1-3 phrases). Modalities other than declarative are at present synthesized by superimposition of artificial prosody (see 5.3). The phonetically dense sentences were extracted by a greedy algorithm from a set of 12,000 well formed Italian sentences, subset of a larger corpus of 200,000 sentences from books and newspapers, previously analyzed to compute the reference coverage of the language in terms of positional diphones and syllables. Acoustic specialization has been realized for a train information service (430 sentences added to the standard database, including foreign phonemes) and a street navigation system (250 sentences). In both cases the sentences were designed starting from a corpus of messages representative of the application. The resulting synthesis is very close to pre-recorded speech quality.

6.2 Tuning of Selection Criteria

Evaluation of various combinations of selection criteria (see 5.1) and tuning of the relevant weights was carried out, to couple score maximization with synthesis quality. For declarative modality (the one represented in the current database) high naturalness is obtained by simple waveform concatenation of synthesis units selected with criteria (1), (2), (3), (5), which are then chosen as a default. The first criterion, *phonetic context* (1), reinforced by the *concatenation factor* (5), favours the selection of continuously articulated sequences, preferably cut on consonants, whereas criterion (2) searches for well connected prosodic sequences, reducing pitch jumps and timing irregularities. To ensure a plausible prosody, criterion (3) restricts the choice to units coming from a coherent position inside a suitable phrase type. If criterion (4) is adopted instead of (2) and (3), looking for a match with target prosodic values defined by rules, the result is much more fragmented. A measure of synthesis fragmentation is provided by the average unit size. On a test corpus of more than 7 million phonemes, with the default criteria more than 93% of the corpus is covered by units larger than traditional diphones (see Fig. 1) and about 25% by triphone units (4 demi-phones). Unit size can be sometimes as long as a whole phrase, and rarely longer than 30 demi-phones. On the same corpus, the use of criterion (4) leads to a shorter unit size, raising from 7% to 35% the text percentage covered by diphones, mainly due to prosodic divergences between natural and rule-driven intonation contours.

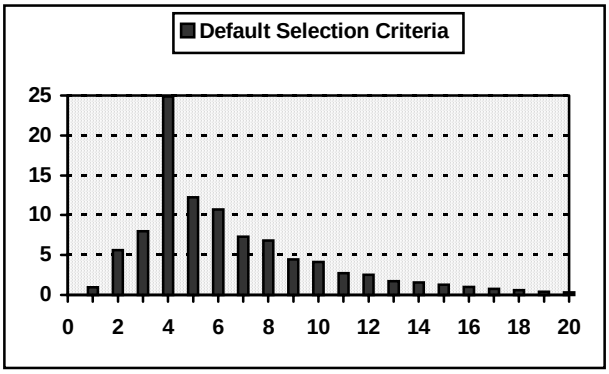

Fig. 1 – Phonetic coverage vs unit size (in demi-phones).

7. CONCLUSION

A corpus-based technique has been applied in the recent evolution of CSELT text-to-speech synthesis. A highly natural-sounding voice is obtained by selecting synthesis units according to phonetic context, prosodic structure and prosodic continuity, and by simply concatenating them, with possible prosodic smoothing at unit junctions. A human-voice effect is ensured by the fact that both timbre and prosody are mostly kept unaltered. In such approach, voice quality is strictly dependent on the contents of the acoustic database and a drawback is the lack of precise prosodic regulation. At present, anomalies in the data might occasionally cause inappropriate contours, while for complete prosodic flexibility the enforcement of artificial prosody is still necessary. Further research will improve prosodic control and develop richer acoustic databases, for new voices and for languages other than Italian.

8. REFERENCES

[1] W.N. Campbell and A.W. Black, "Prosody and the Selection of Source Units for Concatenative Synthesis", in: J. van Santen et al. (eds.), *Progress in Speech Synthesis*, pp. 279-292, Springer New York, 1996.

[2] M. Balestri, S. Lazzaretto, P.L. Salza and S. Sandri, "The CSELT System for Italian Text-to-Speech Synthesis". *Proc. EUROSPEECH '93*, Berlin, Vol. 3, pp.2091-2094.

[3] R. Billi, F. Canavesio, A. Ciaramella and L. Nebbia: "Interactive Voice Technology at Work: the CSELT Experience", *Proceedings of IVTTA '94*, Kyoto.

[4] L. Nebbia, S. Quazza and P.L. Salza, "A Specialised Speech Synthesis Technique for Application to Automatic Reverse Directory Service", *Proceedings of IVTTA '98*, Torino, pp. 223-228.

[5] J.P.H. van Santen and A.L. Buchsbaum, "Methods for Optimal Text Selection", *Proceedings of EUROSPEECH '97*, Rhodes, Vol. 2, pp. 553-556.

[6] B. Angelini, C. Barolo, D. Falavigna, M. Omologo and S. Sandri, "Automatic Diphone Extraction for an Italian Text-to-Speech Synthesis System", *Proceedings of EUROSPEECH '97*, Rhodes, Vol. 2, pp. 581-584.

[7] S. Quazza, P.L. Salza, S. Sandri and A. Spini, "Prosodic Control in a Text-to-Speech System for Italian", *Proc. ESCA Workshop on Prosody*, Lund, 1993, pp. 78-81.