



Design of an optimal continuous speech database for text-to-speech synthesis considered as a Set Covering Problem

Hélène François, Olivier Boëffard

IRISA - Institut de Recherche en Informatique et Systèmes Aléatoires
Université de Rennes 1, ENSSAT, Lannion, France
francois@enssat.fr, boeffard@enssat.fr

Abstract

Text-to-speech synthesis can be carried out by concatenation of acoustic units obtained from a continuous speech database. This paper presents the optimization of such a database according to phonetic criteria. A large corpus of texts is assembled (311 572 sentences), phonetized automatically and condensed (12 217 sentences) to retain only 10 tokens of the most frequent triphonemes. This is a NP-hard problem of set covering. It has been solved in an approximate way using a greedy algorithm. The condensed database covers 25% of the initial distinct triphonemes, each being represented by 10 tokens at least, which allows 95% of the triphoneme tokens of the initial corpus to be covered. The distribution of the triphonemes remains proportional to their initial statistical appearance.

1. Introduction

State-of-the-Art in text-to-speech synthesis consists in juxtaposing pre-recorded acoustic units, typically phones, diphones or polyphones. The multiple representation of these units at acoustic level and in a variety of syntactico-prosodic configurations enables voice quality to be improved significantly [1] [2]. A more radical approach, proposed by R. E. Donovan [3], consists in drawing the sound matter from a continuous speech database. The concept of a unit is then called into question in that the acoustic segments no longer have any pre-defined linguistic limits.

The work we present fits into the general framework of text-to-speech synthesis systems using a continuous speech database. The linguistic units useful to the synthesis are defined a posteriori, i.e. after knowledge of the phonetic string to be synthesized. It is therefore essential that the database of continuous speech should cover as well as possible most of the acoustic sequences used in a language. The aim is then to capture a maximum amount of coarticulation in units longer than conventional diphones or triphones. It would indeed be detrimental for the quality of the voice synthesis if a frequent unit was missing, for example.

The speech database can be built at random [3] or optimized in its *acoustic* [4] or *textual* state; we propose a solution making it possible to automatically constitute a set of textual sentences to be recorded which have acoustic unit covering properties which are satisfactory for new generation speech synthesis systems. Such a work has been carried out in LIMSI [5] in order to evaluate speech recognition systems. J. P. H. van Santen looked at different variants of an optimization method using "greedy" algorithms with objectives similar to those pursued in

our study [6] [7]. So did O. Boëffard and F. Emerard [8] but they used genetic algorithms and their aim was to optimize the learning database of prosodic models. Their conclusions are clearly in favour of an optimization of the speech corpus.

What is at stake in the work presented here is the design, construction and characterization of a continuous speech database optimized in terms of acoustic criteria. Three stages are required: the creation of an optimal textual database, its recording and the annotation of the acoustic database thus obtained. Only the first step is detailed here; it consists, to begin with, in collecting a large quantity of targeted texts and formatting them (in §2), secondly in phonetizing them and analysing them (in §3) and finally in condensing them according to phonetic criteria; the principle of condensation, which is similar to a set covering problem, and its solution, which uses "greedy" algorithms, are presented in §4 and the results set out in §5.

2. Constitution of a large textual corpus

This section presents the methodology we followed to constitute a textual corpus phonetically representative of the language. It is called the "Irisa" corpus. The phonetic richness of a textual body is here measured in the number of triphonemes which are found in one of its most probable phonetic transcriptions. The *triphoneme* is here defined as a sequence of three phonemes and presented as the phonetic description of an allophone. We distinguish between *distinct* triphonemes and triphonemes *tokens*; the latter are individual instantiations of the former. In view of the future extension of this work at a syntactic and prosodic level, our corpus must also contain varied prosodic configurations, which forced us to consider the sentence to be a indivisible linguistic unit.

2.1. Texts selection

We make the assumption that the phonetic richness of a textual corpus depends on that of its lexicon, and that its prosodic richness is in correlation with its syntactic variety, with its stylistic diversity and with its level of language.

Below are presented the 7 databases of the Irisa corpus characterized according to their lexicon (thematic, varied or rich in greek and latin roots), to their syntax (restricted to the length of sentences), to their style (narrative or speech) and to their level of language (familiar, standard, elevated or conventional):

- 39 interviews of comic strip authors (thematic lexicon, speech, familiar & standard languages)
- 50 scripts translated from the american television serial "Friends" (thematic lexicon, short sentences, speech, familiar language)
- the whole book "Jacques le fataliste et son Maître" writ-

This work is granted by France Telecom R&D within contract with DIH/ISP/VMI laboratory.



- ten by Denis Diderot (thematic lexicon, narrative & speech, elevated language)
- 471 Medicine handouts of Claude Bernard Lyon 1 University (rich in roots lexicon, long sentences, narrative, conventional language)
- 1671 articles of “Le Monde” newspaper (varied lexicon, long sentences, narrative, elevated language)
- 3757 Multimanía forums messages (varied lexicon and syntax, narrative and speech, familiar and standard language)
- 95 transcriptions of the French National Assembly sessions (varied lexicon, long sentences, narrative & speech, conventional language)

Irisa corpus contains therefore a varied lexicon; Medicine brings many greek and latin roots. It includes more speech than narrative; all the language levels are represented.

2.2. Texts cleaning and formatting

Texts source language is source HTML which is richer than standard textual formats; it allows for example automatic removal of bold printed speaker names in some dialogues. Texts are cleared of any malformed symbol and converted in a unique paragraph-based textual format. Then they are ready to be phonetized. These preprocessings are quite similar to those applied to BREF corpus [5].

3. Phonetization

The condensation of the Irisa corpus requires its phonetization. The phoneme sequences are determined automatically by a phonetizer from the text developed by France Telecom R&D [9]. Each paragraph is divided into sentences, then into breath groups according to major punctuation boundaries, and translated into a single phonetic string. Phonetization being automatic, there will necessarily be differences between the phonetic string delivered and the string produced by the speaker at the recording of the optimal base. The recording of the whole corpus would enable this to be avoided, as phonetization could be carried out from acoustic production, but such a recording cannot be envisaged. We make the assumption that those variations will only be allophonic.

In this section three corpora are compared: the Irisa corpus, the BREF corpus described in §1 and the Tubach and Boë corpus [10]. The latter has got 304 752 phones which have been transcribed manually.

3.1. Characterization of Irisa corpus

TAB. 1 – Characterization of corpora.

Number	Irisa	Bref	Tubach
Sentences	311 572	167 359	–
Phones	22 960 649	16 416 738	304 752
Phones per Sentence	73,7	98,1	–
Phonemes N_{ph}	34+1	34+1	34
Triphonemes	27 489	25 999	12 019
% N_{ph}^3	64,1	60,6	30,6

Tubach & Boë corpus contains 33 or 34 distinct phonemes depending on the schwa which hasn't always been phonetized. The 2 other corpora include these 34 phonemes and take as well *pause* into account, which explains the “34+1” notation in table

1. The cube N_{ph}^3 of the phonemes number N_{ph} has no linguistic meaning but gives an idea of triphonemes number bound.

The average of the number of phones per sentence is greater in BREF corpus than in the Irisa one (98.1 vs 73.7 phones/sentence). This is due to the chosen texts; BREF corpus contains the long sentences of “Le Monde” newspaper while Irisa corpus includes many dialogue sentences which are shorter.

The 40% increase (16 416 738 to 22 960 649 phones) between BREF and Irisa corpora brings about a only 5% increase of the triphonemes number. Thus we can be optimistic as for the ability of these 2 corpora to phonetically represent French language.

3.2. Phonemes and triphonemes distributions

Let $f(\cdot)$ be the appearance frequency of any phoneme. It goes from 0.2 to 8% of the tokens in Irisa and Tubach & Boë corpora. Let $\Delta(\cdot) = f(\cdot)^{Tubach} - f(\cdot)^{Irisa}$ be the difference of appearance of a phoneme between the 2 corpora. Since 5 phonemes have not been phonetized in the same way in the 2 corpora ([e], [ə], [ø], [œ] and pause), following values were calculated with the 30 remaining phonemes. So, the mean of $\Delta(\cdot)$ is 0.08 which is due to the difference of total distinct phonemes number; its standard deviation is 0.35 then phonemes distributions of the 2 corpora remain comparable despite their different sizes (Irisa corpus is 75 times bigger than Tubach & Boë corpus) and despite their different ways of phonetization (automatic for Irisa and by hand for Tubach & Boë).

Triphonemes distribution is shown on figure 1. It is of type logarithmic; many triphonemes are few represented (2 378 have got only 1 token) while many others have got many tokens (the most frequent triphoneme has got 150 455 tokens). This distribution allows the corpus condensation because of its logarithmic type; so the aim is to make it linear.

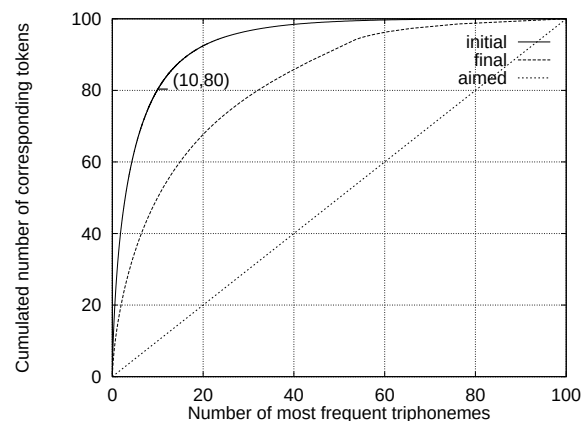


FIG. 1 – Triphonemes distribution of Irisa databases. 10% of the most frequent distinct triphonemes cover 80% of the triphonemes tokens.

3.3. Database type influence

We noted that the ratio of the triphonemes common to n databases decreases exponentially with n . Thus with the databases defined in §2.1 this ratio goes from 60% with 2 databases to 32% with the 7 databases. But these 32% allows the covering of 95% of the triphonemes tokens of the whole Irisa corpus.



4. Textual corpus condensation

Sections 3.2 and 3.3 show that corpus can be condensed by minimizing the tokens number of frequent units and by removing units which are rare, i.e. few representative. That can be considered as a Set Covering Problem (SCP).

4.1. Minimal Set Covering Problem

4.1.1. Problem formalization

We adopt the following notations :

- Let $\mathcal{A}$ be a finite set.
- Let $\mathcal{F}$ be a collection of subsets S of $\mathcal{A}$.
- A cover $\mathcal{C}$ of $\mathcal{A}$ is a subset of $\mathcal{F}$ in conformity with equation 1.
- Let w be a function linking a cost with a given cover of $\mathcal{A}$. The cover of minimum cost associated with function w is defined according to equation 2.

$$\bigcup_{S \in \mathcal{C}} S = \mathcal{A}, \quad \mathcal{C} \subseteq \mathcal{F} \quad (1)$$

$$\mathcal{C}^* = \arg \min_{\mathcal{C} \subseteq \mathcal{F}} w(\mathcal{C}) \quad (2)$$

4.1.2. Representation format

We adopt the following representation format :

- $\mathcal{A}$ is represented by a row vector; $\mathcal{A} = \{a_j\}$, $j \in N$, $N = \{1, \dots, N\}$.
- $\mathcal{F}$ is represented by a column vector; $\mathcal{S} = \{S_i\}$, $i \in P$, $P = \{1, \dots, P\}$.
- Let $\mathcal{M} = \{m_{ij}\}$, $i \in P$, $j \in N$ be the matrix representing $\mathcal{F} \times \mathcal{A}$; a row $S_i \in \mathcal{F}$ covers a column $a_j \in \mathcal{A}$, or a column a_j belongs to S_i if $m_{ij} > 0$.
- Let $\mathcal{W}$ be the column vector representing the cost of lines; $\mathcal{W} = \{w_i\}$, $w_i > 0$ where w_i is the cost of line S_i , $i \in P$.

A cover of $\mathcal{A}$ is a subset $\mathcal{C}$ of lines so that each column a_j of $\mathcal{A}$ is covered by at least one line of $\mathcal{C}$ as shown in equation 3, where $x_i \in \{0, 1\} \quad \forall i \in P$, and $x_i = 1$ if $S_i \in \mathcal{C}$, 0 otherwise.

A cover $\mathcal{C}^*$ is minimal if its cost $w(\mathcal{C}^*)$ is a minimum of the cost function defined by equation 4.

$$\mathcal{C} \subseteq \mathcal{F} \quad \Longleftrightarrow \quad \forall j \in P, \quad \sum_{i \in N} m_{ij} x_i \geq 1 \quad (3)$$

$$w(\mathcal{C}^*) = \sum_{i \in P} w_i x_i \quad (4)$$

4.1.3. Exact and approximative solutions

This problem has an exact solution. In theory one just have to take into account the set of all possible subsets of $\mathcal{F}$, to restrict it to the subsets which are covers of $\mathcal{A}$, to calculate the corresponding costs and to extract the cover(s) of minimum cost. This solution is redoubtable as regards combinatory considerations. Research has been carried out to achieve this exact solution with a lower calculation cost, notably by Beasley & Jörnsten [11]. Nevertheless resulting algorithms can only be used with about few thousands of rows and few hundreds of columns.

Approximative solutions are thus inescapable; they resort to genetic algorithms [8] or to Lagrangian relaxation [12]. We present in this work a solution using a greedy algorithm [13].

4.1.4. Basic greedy algorithm

In the following greedy algorithm, operator “+” corresponds to the addition of an element in a set, and $S_i \setminus S_j$ means S_i deprived of its intersection with S_j .

```

·  $\mathcal{C} = \emptyset$ 
· while  $(\bigcup_{S_i \in \mathcal{C}} \{S_i\} \neq \mathcal{A})$  do
  · Choice of an optimal  $S_j \in \mathcal{F}$ 
  ·  $\mathcal{C} = \mathcal{C} + S_j$ 
  ·  $\forall S_i \in \mathcal{F}, S_i = S_i \setminus S_j$ 
  ·  $\mathcal{A} = \mathcal{A} \setminus S_j$ 
· end

```

The major difficulty is the choice of S_j ; indeed optimum can be calculated according varied criteria.

4.2. Application to linguistic units

4.2.1. Covering triphonemes with a minimum set of sentences

In this work, $\mathcal{F}$ is the set of Irisa corpus sentences (composed of triphonemes) and $\mathcal{A} = \bigcup_{S_i \in \mathcal{F}} \{S_i\}$ is the set of corresponding distinct triphonemes. The aim is to extract a set $\mathcal{C} \subseteq \mathcal{F}$ of sentences enabling the cover of all the units of $\mathcal{A}$, this set having a minimum number of units. The way of building $\mathcal{A}$ ensured its ability of being totally covered, for example by $\mathcal{F}$.

4.2.2. Additional constraint R_{min}

The aim is not only to cover the triphonemes. We need to have at least $R_{min} \simeq 10$ tokens of each too. This is required for postprocessings consisting in learning acoustic models. This constraint means that total cover is then impossible because several units have less than R_{min} tokens in the whole corpus. Two solutions are conceivable; the first one is to proceed to a partial covering, and the second one is to lead the problem back to a total set covering problem. We chose the second solution, the first one forbidding the use of some specific methods (focusing on rare units for example).

A preprocessing has thus been carried out consisting in restricting set $\mathcal{A}$ to units having at least I_{min} tokens in the corpus. One must have $I_{min} \geq R_{min}$ for the total covering to be possible. R_{min} and I_{min} values are debated in §5.1.

4.2.3. Greedy algorithm criteria

Each iteration of the greedy algorithm requires the choice of a sentence which can be considered optimal if it possesses for example:

- a maximum of distinct units,
- a maximum of units tokens,
- the rarest unit,
- a combination of the above criteria possibly weighed by the sentence length.

The choice of a criterion depends on the covering type. Indeed the first two criteria have very different results with phonemes while they are almost equivalent with triphonemes.

5. Results

5.1. Preprocessing

The consequences of I_{min} choice (see §4.2.2) are shown in figure 2 which is limited to the units having less than 2000 tokens (the whole curve goes up to 150 455 tokens). If triphonemes with less than 10 tokens are removed from $\mathcal{A}$, only 70% of the distinct triphonemes are kept; but they make it possible to cover 99.9% of the corpus tokens. 70% corresponds 19 508



distinct triphonemes. If we consider 10 tokens for each, an ideal covering, and an average duration of $80ms$ per associated allophone, it represents $19508 \times 10 \times 80ms = 4,3$ hours of speech. That's why preprocessing is done with $I_{min} > R_{min}$ in fact. A compromise between the number of well-represented units and recording duration must be achieved. Besides, the greater is I_{min} , the better is the covering because the greater is the choice of R_{min} units between at least I_{min} tokens.

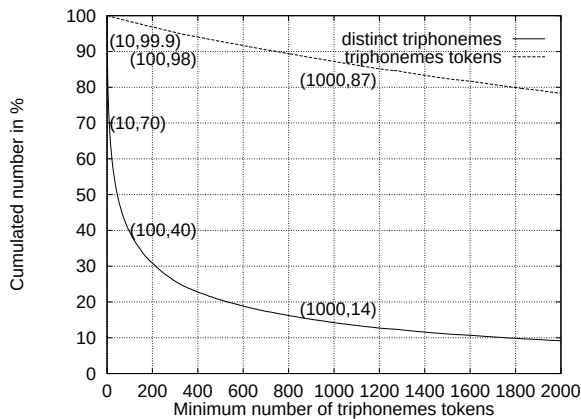


FIG. 2 – Losses caused by preprocessing. After removal of distinct triphonemes having less than 1000 tokens, 14% live on which can cover 87% of corpus tokens.

5.2. Set covering with greedy algorithm

TAB. 2 – Characterization of initial and condensed databases.

Number	initial	condensed
Sentences	311 572	12 217
Distinct triphonemes	27 489	13 102
with 10+ tokens	19508	6 982
% of 27489	71%	25%
Triphonemes tokens	22 337 505	217 593
Phonemes/Sentence	73,7	19,8

Table 2 shows the differences between initial corpus and condensed database. Condensation is intense; only 0.97% of the corpus tokens have been kept. 25% of the distinct triphonemes live on; they allow the covering of 95% of the tokens. Therefore 1 token every 20 will be missing. The average number of phones per sentence fell down which means that short sentences have been privileged; indeed dialogue sentences have essentially been retained. Figure 1 shows triphonemes distributions before and after condensation; it is more linear than before i.e. the number of tokens per unit is more balanced.

6. Conclusions

This paper presents the optimisation, according to phonetic criteria, of a textual database of continuous speech, by condensation of a broad textual corpus. This Set Covering Problem has been solved using greedy algorithms.

The objective was to attain a uniform distribution of the most frequent triphonemes. It is indeed more balanced than in the initial corpus, but remains proportional to the initial statis-

tical appearance of the triphonemes. This is due to the non-optimality of the algorithms used. The most frequent triphonemes are therefore over-represented in the condensed database; this can prove to be very useful for the future units selection, but can only be evaluated at that moment. The optimal database represents 1% of the Irsa corpus; with 25% of the initial distinct triphonemes, it can cover 95% of the initial tokens.

This work can be extended by optimising the tokens of a triphoneme, so that they have left and right contexts which are both diversified and common. At the moment we study such an approach [14].

7. References

- [1] A. Black and P. Taylor, "The Festival speech synthesis system" System documentation HCRC/RT-83, University of Edinburgh, UK, 1997.
- [2] M. Beutnagel et al., "The AT&T Next-gen TTS System", In Proc. of the Joint Meeting of ASA, EAA, and DAGA, Berlin, Germany, 1999.
- [3] R. E. Donovan, "Trainable Speech Synthesis", PHD Thesis, Cambridge University Engineering Department, 1996.
- [4] W. N. Campbell and A. W. Black, "Prosody and the selection of source units for concatenative synthesis", In J. van Santen et al., editors, *Progress in Speech Synthesis*, Springer Verlag, 1996.
- [5] J.-L. Gauvain et al., "Design considerations and text selection for bref, a large french read-speech corpus", In Proc. of ICSLP, Vol. 2, pp. 1097–2000, Kobe, Japan, 1990.
- [6] J. P. H. van Santen and A. L. Buchsbaum, "Methods for optimal text selection", In Proc. of Eurospeech, pp. 553–556, Rhodes, Greece, 1997.
- [7] J. P. H. van Santen, "Combinatorial issues in text-to-speech synthesis", In Proc. of Eurospeech, pages 2511–2514, Rhodes, Greece, Sep 1997.
- [8] O. Boëffard and F. Emerard, "Application-dependent prosodic models for text-to-speech synthesis and automatic design of learning database corpus using genetic algorithm", In Proc. of Eurospeech, Rhodes, Greece, Sep 1997.
- [9] D. Larreur et al., "Linguistic and prosodic processing for a text-to-speech synthesis system", In Proc. of Eurospeech, Vol. 2, pp. 510–513, Paris, Sep 1989.
- [10] J.-P. Tubach and L.-J. Boë, "Un corpus de transcriptions phonétiques: constitution et exploitation statistique", Rapport ENST - 85D001, Ecole Nationale Supérieure des Télécommunications, Paris, Avril 1985.
- [11] J. E. Beasley and K. Jörnsten, "Enhancing an algorithm for set covering problems", *European Journal of Operational Research* 58 pp. 293–300, June 1992.
- [12] A. Caprara et al., "A heuristic method for the set covering problem", *Operations Research*, Vol. 47, No 5, pp. 730–743, Sep-Oct 1999.
- [13] V. T. Paschos, "A Survey of Approximately Optimal Solutions to Some Covering and Packing Problems", In *ACM Computing Surveys*, Vol. 29, No. 2, June 1997.
- [14] O. Boëffard, "Variable-length acoustic units inference for text-to-speech synthesis", Submitted to Eurospeech, Aalborg, Denmark, 2001.