Towards optimal TTS corpora

Didier Cadic [1], Cédric Boidin [1], Christophe d'Alessandro [2]
[1] Orange Labs, France
[2] LIMSI, France
E-mail: {didier.cadic, cedric.boidin}@orange-ftgroup.com, cda@limsi.fr

Abstract

Unit selection text-to-speech systems currently produce very natural synthesized phrases by concatenating speech segments from a large database. Recently, increasing demand for designing high quality voices with less data has created need for further optimization of the textual corpus recorded by the speaker. This corpus is traditionally the result of a condensation process: sentences are selected from a reference corpus, using an optimization algorithm (generally greedy) guided by the coverage rate of classic units (diphones, triphones, words…). Such an approach is, however, strongly constrained by the finite content of the reference corpus, providing limited language possibilities. To gain flexibility in the optimization process, in this paper, we introduce a new corpus building procedure based on sentence construction rather than sentence selection. Sentences are generated using Finite State Transducers, assisted by a human operator and guided by a new frequency-weighted coverage criterion based on Vocalic Sandwiches. This semi-automatic process requires time-consuming human intervention but seems to give access to much denser corpora, with a density increase of 30 to 40% for a given coverage rate.

1. Introduction

During the last 15 years, the emergence of corpus-based concatenative speech synthesis systems has given rise to major improvements in Text-to-Speech (TTS). Their success relies mainly on the use of large speech databases containing several hours of recordings of a single speaker. Recently, increasing demand for designing high quality voices with less data has created need for further optimization of the speech databases.

The recorded script (or corpus) is expected to provide a wide variety of phonetic and prosodic events for a minimal set of sentences. It is traditionally the result of an optimization process, which raises the two following problems: 1- Which phonetic and prosodic criteria are best suited for this stage? 2- Which optimization algorithm should be used?

In section 2, we describe a novel approach to the first problem, focused on concatenative TTS specific features. In section 3, we introduce a new corpus building procedure based on sentence construction rather than sentence selection. In section 4, we discuss benefits, drawbacks and perspectives of this building procedure. In the last section, we summarize our contributions and present future work.

2. Corpus optimization approach

2.1 Concatenative TTS framework

At the starting point of a concatenative TTS system, a large textual corpus is read by a speaker, resulting in a speech database that is typically several hours long.

Then, to vocalize an input text, the system automatically selects and concatenates units from the database (Sagisaka, 1988; Hunt & Black, 1996). The selection step aims at minimizing a cost function which is traditionally composed of a target cost and a concatenation cost. The former gives, for each candidate unit, a distance between its original context (prosodic, linguistic) and the target one. The latter measures the smoothness of the concatenation between consecutive candidate units. The unit sequence presenting minimal cost is then supposed to exhibit relevant prosody and smooth concatenations. Both costs may integrate symbolic features (phonetic/linguistic context…) as well as acoustic features (pitch, duration, spectrum…) of recorded speech.

The initial textual corpus, called *synthesis corpus*, can also be seen as the result of a preliminary optimization process. It is supposed to statistically maximize the perceptual quality of the final TTS system.

2.2 Traditional optimization criterion

Most state-of-the-art synthesis corpora are designed to maximize coverage rates of traditional units : diphones, triphones or even quadriphones, words, *etc*., often enhanced with contextual information (Black & Lenzo, 2001).

Such well-known units are however not dedicated to concatenative speech synthesis. They are of general use in speech technologies and linguistics (Gauvain et al, 1990).

2.3 Connection with perceptual quality

In order to optimize the synthesis corpus in terms of final speech quality, we suggest relying on the automatic and efficient cost function of the selection process, designed to quantify human perception. However at the stage of corpus design, acoustic features are not available (the speaker is waiting for the script!). Only the symbolic part of the cost function can be computed.

The concatenation cost, usually based on acoustic distances, can be satisfyingly projected on symbolic features. Indeed acoustic mismatches responsible for concatenation artefacts show high dependency on the phoneme type (Yi & Glass, 1998). Such mismatches tend

to be more audible on phonemes presenting:

- high context-dependency, like liquids, vowels and semi-vowels, since coarticulation effects may result in significant inter-occurrences spectral variability (Lindblom, 1963)
- high spectral stability, like vowels, where discontinuities are scarcely acceptable
- high energy, for obvious perceptual reasons
- voicing, because of periodicity breaks
- large vocal tract opening, since such configuration results in more acute formants and therefore requires precise formantic continuity.

Then a simple criterion consists in penalizing concatenations depending on the phoneme type on which they occur. For instance, in decreasing penalization order, vowel > semi-vowel > liquid > consonant.

Such symbolic approximation of the concatenation cost leads to the definition of a new unit, the **vocalic sandwich**, which we introduced previously (Cadic et al, 2009). This unit represents any sequence of "fragile" phonemes (like vowels and semi-vowels where high concatenation costs are generally observed), surrounded by two "robust" phonemes (typically consonants). Table 1 shows an example of a French sentence and its decomposition into vocalic sandwiches and consonant clusters. One can notice that some sandwiches extend across word boundaries. For further details please refer to (Cadic et al, 2009).

(1)	*Et ce week-end sera exceptionnel.*
(2)	# e s ə w i k ɛ n d s ə ʁ a ɛ k s ɛ p s j ɔ n ɛ l #
(3)	#e s k ɛ n ʁ a ɛ k s j ɔ n s ə w i k s ə ʁ s ɛ p n ɛ l
(4)	n d s k s p s l #

Table 1: example of a French sentence (1), along with its phonetic transcription (2) and its split into vocalic sandwiches (3) and consonant clusters (4).
Translation: "*And this week-end will be exceptional.*"

In a way, for a given sentence to synthesize, the proportion of vocalic sandwiches found in the synthesis corpus is related to the proportion of joins that the TTS system is able to place on consonants. The link with the concatenation cost is thus straightforward. Furthermore, in their context-dependent version, **vocalic sandwiches are enhanced with symbolic information about linguistic and prosodic contexts**, similarly to the target cost.

Therefore the Vocalic Sandwiches Coverage Rate (VSCR) can be seen as an analogous of the selection cost in terms of coverage rate, which is a much easier concept to handle. For information, using the Orange Labs (ex- France Telecom R&D) diphone-based TTS system, VSCR shows a correlation of -0.77 with the selection cost, and 0.46 with human perception, in comparison to respectively -0.52 and 0.18 for context-dependent diphones (Boidin et al, 2009).

2.4 Choice of a reference corpus

When designing a synthesis corpus, the VSCR is computed in regard to a *reference corpus*. A sandwich representing n occurrences upon a total of N sandwiches in the reference corpus, increases the VSCR by n/N the first time it is introduced in the synthesis corpus.

We used a French reference corpus of approximately 2,500,000 words, consisting of SMS, recent books, contemporary theatre plays, newspaper articles, films and series subtitles, vocal server messages, recipes, newsgroup posts and instant messages. Spelling and phonetic transcription were partially reviewed. In this corpus we collected 3,600,000 context-dependent sandwiches, of which only 93,000 were distinct. The 5,400 most frequent sandwiches covered 80% of the reference corpus.

It is important to note that such a frequency-driven approach, motivated by our analysis in paragraph 2.3, is not yet universally accepted in TTS (van Santen, 1997). Many studies focus on set-covering strategies where target units are determined a priori (van Santen & Buchsbaum, 1997; François & Boëffard, 2001).

We hereafter restrict the notion of "sentences" to single breath-groups, to increase flexibility in the corpus constitution process, while limiting variability and errors at the recording stage.

2.5 Greedy optimization

The VSCR optimization over all possible synthesis corpora has been proven to be an NP-hard problem (Garey & Johnson, 1979). Instead of searching for a global optimum, which may be computationally out or reach, we adopted a classical "greedy" approach (van Santen & Buchsbaum, 1997). Sentences were incrementally added to the synthesis corpus, by maximizing at each step the increase in VSCR. The source of these sentences will be discussed in the next section.

The result of such greedy optimization is of course suboptimal but offers advantages. First it is computationally very efficient, then it guarantees some kind of optimality even if only the beginning of the synthesis corpus is recorded. Indeed the amount of recordings targeted for a voice creation highly depends on needs and budget. Therefore a **scalable synthesis corpus**, allowing partial recording with moderate loss of quality, is attractive.

3. Constitution of the synthesis corpus

3.1 Optimal distribution

The cumulative distribution function of all sandwiches occurrences collected in the reference corpus shows the optimal coverage rate we could reach in a synthesis corpus (upper curve in Figure 1). This optimum corresponds to a compact corpus containing only one occurrence of the most frequent sandwiches. However such density is out of reach since language perplexity imposes at least some dispersion and redundancy. In the

next sections we will try to make up a synthesis corpus closest to this optimum, while ensuring correctness and readability of the sentences.

3.2 Corpus condensation

One way of constituting such corpus is to iteratively select in a "pick corpus" sentences maximizing the increase in VSCR. Traditionally pick and reference corpora are identical. Such **corpus condensation** has been extensively used with different criteria (François & Boëffard, 2002).

We observed the VSCR evolution throughout the greedy selection process, and compared it with the random selection of sentences as lower bound (Figure 1).

Compared to a random corpus, the distribution obtained when condensing the reference corpus through greedy selection is much denser. But still, the corpus size is about twice the optimal size for a given coverage rate.

According to (Chevelu et al, 2007), performance of the greedy selection algorithm is close to that of an optimal condensation algorithm, with only 10% sub-optimality in corpus size. It can therefore be concluded that all condensation approaches are strongly constrained by the limited combinations encountered in the finite pick corpus.

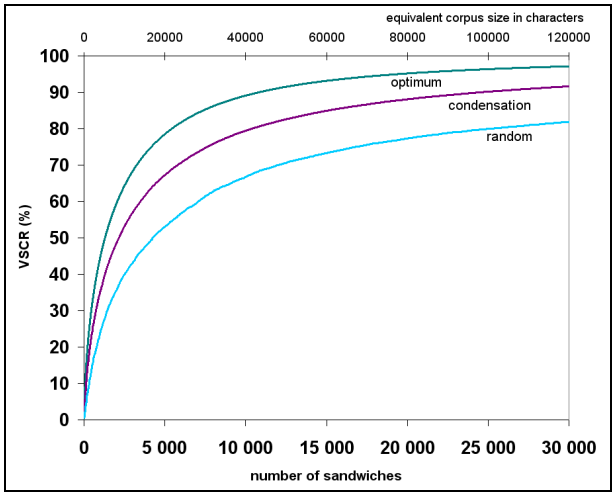

Figure 1: VSCR evolution with a greedy condensation process, compared to random and optimum.

3.3 Sentence construction

We explore here a new approach of the synthesis corpus constitution. To gain flexibility in the greedy process, we suggest building sentences instead of selecting them from a pick/reference corpus.

3.3.1 Semi-automatic corpus building algorithm

Starting from an initial corpus, the proposed algorithm computes a new optimal sentence, composed of the **sandwich sequence** that maximizes the increase in coverage rate.

Sentences are generated using finite state transducers (FST), handled with OpenFst (Allauzen et al, 2007). States correspond to sandwiches, and arcs to allowed transitions, i.e. transitions between successive sandwiches that were observed in the reference corpus (sandwich 2-grams). Arc costs are set to $(1-\delta_{VSCR}) \in [0; 1]$, where δ_{VSCR} denotes the increase in VSCR brought by the ending sandwich of the arc. Costs are updated after each sentence construction (put to zero when the ending sandwich is covered).

Optimal sandwich sequences can be (nearly) instantaneously obtained by bestpath searches. However, simple cost minimization leads inevitably to very short solutions. Path costs should therefore be averaged over their length. In order to do this we changed the topology of our FST (through graph composition) so that it forces the length of paths (sandwich sequences going from initial to final "silence" states) to a fixed value. We kept 15 FST versions, one for each possible sentence size (with a maximum of 15 sandwiches). Then, the optimization step consists in 15 parallel bestpath searches, allowing easy management of size effects (minimization of mean cost instead of total cost, limitation of sentence size and infinite loops…).

Since there is neither syntactic nor semantic consideration in the FST, generated sequences are likely to be nonsense, and even not lexically correct (not made of French words). But in practice, the 2-grams constraint imposes some medium-term coherence to the sequences, thanks to an average length of 5.1 phonemes and the contextual information attached. Therefore, one can always identify large sub-sequences that may be part of a rational sentence. However this approach cannot be fully automated, as human linguistic expertise is required.

A dedicated tool was developed for semi-automatic sentence building. Considering an initial sequence automatically generated through mean-cost minimization, the operator has to identify a "promising" sub-sequence, *i.e.* carrying embryonic sense, and then ask for the generation of a new environment. More precisely, the beginning (or the ending) of the computed optimal sequence being defined, the operator can ask openFST for the second bestpath ending (or beginning). This way, the operator is always guided towards the most frequent and uncovered sandwiches and can iteratively build an acceptable and almost optimal sentence (Figure 2) which he finally transcripts verbatim. The operator has to be familiar with phonetics. Nevertheless TTS can be run at each step to help to lexically interpret the sequences. Several other functionalities are available: *"Cancel"*, *"Reset"*, *"Force quick sentence ending"* and *"Force quick sentence beginning"* (for situations where a short sub-sequence is almost satisfying but environment regenerations return sequences that are too large).

This procedure is relatively time consuming: 3 minutes (around 50 steps) on average to build a plausible sentence. This is a major drawback of the process, but it must be emphasized that the synthesis corpus needs to be constructed only once, and can be reused for many voice creations. More constrained generation models could certainly reduce this time or even avoid human intervention, but the loss of flexibility could result in lower density.

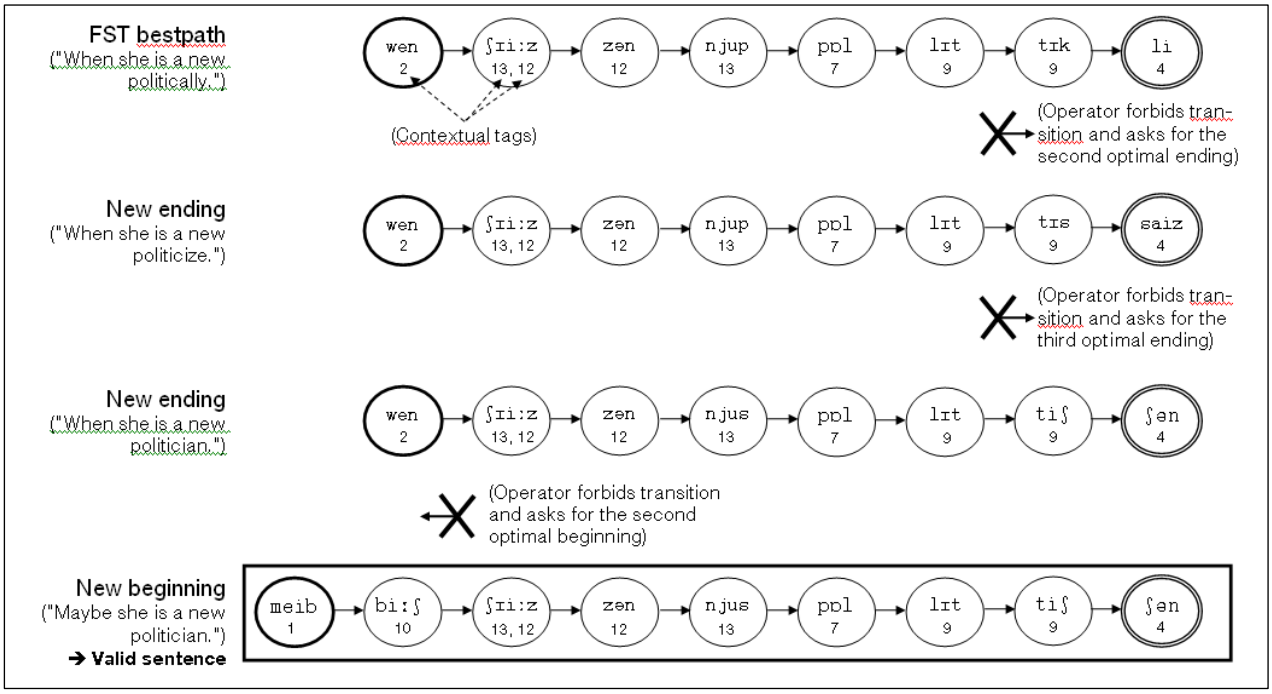

Figure 2: Virtual and simplistic example of the sentence construction process, applied on a short sentence.

3.3.2 Coverage potential estimation

We did not apply this semi-automatic process on a full corpus yet. Beforehand we did a rough estimation of its coverage potential.

An upper bound (named "*auto_construction*") was computed by automatically validating the bestpath at each step of the greedy process, thus building incorrect but FST-optimal sentences.

For the lower bound, the semi-automatic building process was used for small groups of 10 to 30 sentences, **alternately** with bigger sessions of automatic construction. This way we obtained a representative sampling ("*auto_manual_construction*") of the semi-automatic process, distributed on the entire VSCR scale. The VSCR derivative observed on these samples can give us an idea on what could be the VSCR evolution throughout an entire corpus construction. Indeed, sampled derivative measures of the semi-automatic construction process appeared to match those of a dilated VSCR curve of *auto_construction* (Figure 3). The best correspondence was obtained by expanding by a factor of 1.15 the number of sandwiches in *auto_construction*.

This gives us a lower bound, because of the alternation between automatic and semi-automatic construction. Indeed, automatic sessions tend to build denser sentences, thus leaving less latitude for the following semi-automatic sessions. VSCR derivatives recorded on the semi-automatic sessions should therefore be greater if no automatic session was used to build the previous corpus.

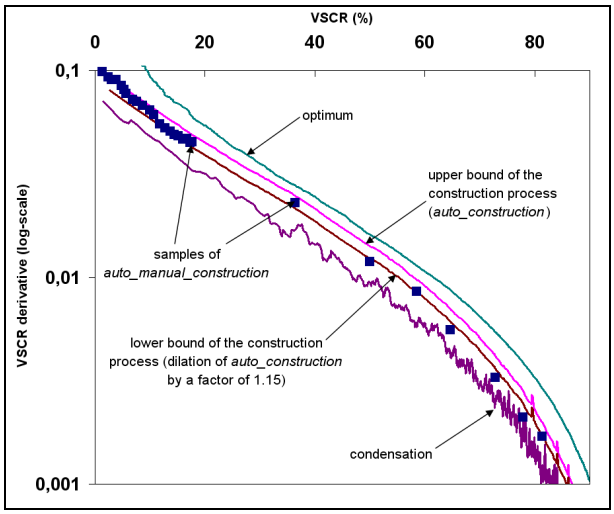

Figure 3: evolution of VSCR derivatives for the different approaches, as a function of the coverage rate. Samples of the *auto_manual_construction* process are well described by a dilation curve of *auto_construction*.

Figure 4 summarizes all observed or estimated distributions. The construction process seems to improve significantly the density of the corpus, compared to a condensation approach. **Constructed corpora are expected to be 30 to 40% smaller** for a given VSCR, which could be very advantageous.

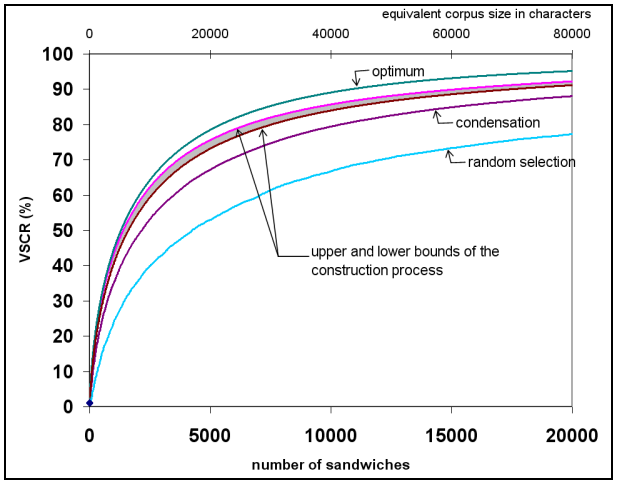

Figure 4: estimation of the VSCR throughout the greedy construction process (grey area), compared to the optimum, the greedy condensation and random selection.

4. Discussion

4.1 About optimality

According to the previous section, our semi-automatic sentence building process shows high potential regarding density and coverage rate.

Following (François & Boëffard, 2002) we could make the final corpus around 10% denser by applying a "spitting" method that involves an *a posteriori* removal of sentences having low impact on the VSCR. This would, however, penalize the scalability of the corpus: its complete recording would then be necessary to take advantage of its optimality (see 2.5).

Either way, our building process seems to outperform the optimal upper bound of selection algorithms given by (Chevelu et al, 2007), although a precise comparison on identical criteria remains to be done.

4.2 About sentence length

In spite of the mean cost computation (see 3.3.1), both selection and construction processes tend to favour very short sentences to gain flexibility in the coverage optimization, at the price however of the reader's convenience and naturalness. With our construction tool, it is possible to manually guide the process towards longer and more "comfortable" sentences. Numeric constraints on the sentence size could also be easily introduced. However the best way to counter this tendency would probably be to enrich linguistic contexts attached to vocalic sandwiches with extended information about the position in the breath group.

4.3 VSCR variants

Applying our semi-automatic sentence building process to the construction of a full synthesis corpus (typically 10-20,000 words for a small corpus) could lead to VSCR above 95%. At this coverage level, each uncovered sandwich occurs less than 10 times in our reference corpus. It could result in acute dependency on the reference sources, which is not desirable. Therefore we suggest using the following variations of our basic VSCR criterion:

- the coverage rate of **sandwich 2-grams with liquids [l] and [ʁ][1] considered as "fragile" phonemes** (*i.e.* they are enclosed in sandwiches, see line 3 of Table 2). Such rich units have an average length of 5.7 phonemes. Their coverage guarantees high quality TTS but is hard to obtain. A reasonable target rate within a 15,000 words corpus could be around 40%, with marginal 2-grams corresponding to 40 occurrences in the reference corpus.
- the coverage rate of **sandwich 1-grams with liquids considered as "fragile" phonemes** (see line 4 of Table 2). These units have an average length of 3.6 phonemes. The target rate could be around 80%, corresponding to 25 occurrences in the reference corpus.
- The coverage rate of "simple" sandwiches, i.e. **1-grams with liquids considered as "robust" phonemes** (see line 5 of Table 2, which is identical to line 3 of Table 1). These basic units have an average length of 3.2 phonemes. At a rate of 90%, marginal sandwiches correspond to 25 occurrences in the reference corpus.

These three criteria of decreasing complexity are illustrations of VSCR variations that could be used consecutively throughout the corpus building process, in order to diversify the coverage of the final corpus in a way favouring TTS segmental quality.

(1)	*Et ce week-end sera exceptionnel.*
(2)	# e s ə w i k ɛ n d s ə ʁ a ɛ k s ɛ p s j ɔ n ɛ l #
(3)	(null-# e s) (# e s-s ə w i k) (s ə w i k-k ɛ n) (k ɛ n-n d s-s ə ʁ a ɛ k) (s ə ʁ a ɛ k-k s-s ɛ p) (s ɛ p-p s-s j ɔ n) (s j ɔ n-n ɛ l) (n ɛ l-#-null)
(4)	#e s k ɛ n s ə w i k s ə ʁ a ɛ k s ɛ p s j ɔ n n ɛ l#
(5)	#e s k ɛ n s ə w i k ʁ a ɛ k s ə ʁ s ɛ p s j ɔ n n ɛ l

Table 2. Sandwich variants (same example as Table 1): sandwich 2-grams (3) and 1-grams (4) with liquids considered as "fragile" phonemes, sandwich 1-grams with liquids considered as "robust" phonemes (5).
Translation: "*And this week-end will be exceptional.*"

4.4 Relevance for real applications

Apart from its density performance, our building process suffers from several drawbacks. First, it requires costly and time-consuming human intervention. Then, we observed that, in general, built sentences have less

[1] IPA notation (International Phonetic Alphabet)

semantic coherence than sentences selected from the reference corpus. This is a logical consequence of our struggle against natural language perplexity: redundancy and dispersion are minimized at the price of semantics. Possible repercussions on the reading stage, like resumptions, unnaturalness, or even irritation of the speaker, could counterbalance density benefits. Further experiments are required to evaluate this point, and a better compromise between density and semantic coherence could be researched, for example by refining our underlying language model.

5. Conclusion

In the framework of corpus-based concatenative TTS, this work focuses on synthesis corpus design.

A new criterion, the vocalic sandwiches coverage rate (VSCR), is introduced as a convenient symbolic approximation of the selection cost. This criterion exhibits higher correlation with the final TTS quality than traditional units. Then a new corpus building technique maximizing this criterion is discussed. Building sentences, instead of selecting them from a reference corpus, seems to give access to much denser corpora, with a density increase of 30 to 40% (Figure 4).

Current works address the generalization of our process on a full synthesis corpus as well as the perceptual evaluation of TTS voices based on various types of corpora: construction vs. selection, VSCR vs. di- or triphones, *etc.*

6. References

Allauzen C., Riley M., Schalkwyk J., Skut W., Mohri M. (2007). OpenFst: A General and Efficient Weighted Finite-State Transducer Library. In *Proceedings of the 9th International Conference on Implementation and Application of Automata (CIAA 2007)*, volume 4783 of *Lecture Notes in Computer Science*, pages 11-23. Springer.

Black A. W., Lenzo K. A. (2001). Optimal data selection for unit selection synthesis. In *proceedings of the 4th ISCA Tutorial and Research Workshop (ITRW) on Speech Synthesis (SSW-4)*, pp. 63-67, Perthshire, Scotland.

Boidin C., Rieser V., Van der Plas L., Lemon O., Chevelu J. (2009). Predicting how it sounds: Re-ranking dialogue prompts based on TTS quality for adaptive Spoken Dialogue Systems. In *proceedings of Interspeech09*, pp. 2487-2490. Brighton, UK.

Cadic D., Boidin C., d'Alessandro C. (2009). Vocalic sandwich, a unit designed for unit selection TTS. In *proceedings of Interspeech09*, pp. 2079-2082. Brighton, UK.

Chevelu J., Barbot N., Boëffard O., Delhay A. (2007). Lagrangian relaxation for optimal corpus design. In *proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW6)*, pp. 211-216, Bonn, Germany.

François H., Boeffard O. (2001). Design of an optimal continuous speech database for text-to-speech synthesis considered as a set covering problem. In *proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pp. 829–832, *Aalborg, Denmark.*

François H., Boeffard O. (2002). The greedy algorithm and its application to the construction of a continuous speech database. In *Proceedings of the 3rd international Language Resources and Evaluation Conference (LREC 2002)*, pp. 1420-1426, Las Palmas, Spain.

Garey M., Johnson D. (1979). Computers and intractability: a guide to the theory of NP-completeness. Freeman.

Gauvain J.-L, Lamel L.F., Eskénazi M. (1990). Design considerations and text selection for BREF, a large French read-speech corpus. In *proceedings of the 1st International Conference on Spoken Language Processsing (ICSLP 90)*, pp. 1097-1100, Kobe, Japan.

Hunt A., Black A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *proceedings of the 21st International Conference on Acoustics, Speech, and Signal Processing (ICASSP96)*, pp. 373-376. Atlanta, USA.

Lindblom B. (1963). Spectrographic study of vowel reduction. In *Journal of the Acoustical Society of America (JASA) 35*, pp. 1773-1781.

Sagisaka Y. (1988), Speech synthesis by rules using an optimal selection of non-uniform synthesis units. In *proceedings of the 13th International Conference on Acoustics, Speech, and Signal Processing (ICASSP88)*, pp. 679-682. New York, USA.

Van Santen J. (1997). Combinatorial issues in text-to-speech synthesis. In *proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, vol. 5, pp. 2511-2514, Rhodes, Greece.

Van Santen J., Buchsbaum A. (1997). Methods for optimal text selection. In *proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech'97)*, pp. 553-556, Rhodes, Greece.

Yi J., Glass J. (1998). Natural-sounding speech synthesis using variable-length units. In *proceedings of the 5th International Conference on Spoken Language Processsing (ICSLP98)*, paper no. 1151. Sydney, Australia.