INTERSPEECH 2009 BRIGHTON

Vocalic sandwich, a unit designed for unit selection TTS

Didier Cadic [1], *Cédric Boidin* [1], *Christophe d'Alessandro* [2]

[1] Orange Labs, France
[2] LIMSI, France

{didier.cadic, cedric.boidin}@orange-ftgroup.com, cda@limsi.fr

Abstract

Unit selection text-to-speech systems currently produce very natural synthetic sentences by concatenating speech segments from a large database. Recently, increasing demand for designing high quality voices with less data creates need for further optimization of the textual corpus recorded by the speaker. The optimization process of this corpus is traditionally guided by the coverage rate of well-known units: triphones, words… Such units are however not dedicated to concatenative speech synthesis; they are of general use in speech technologies and linguistics. In this paper, we describe a new unit which takes account of concatenative TTS own features: the "vocalic sandwich." Both an objective and a perceptual evaluation tend to show that vocalic sandwiches are appropriate units for corpus design.

Index Terms: speech synthesis, concatenation cost, corpus design, vocalic sandwich

10.21437/Interspeech.2009-596

1. Introduction

The emergence of corpus-based concatenative speech synthesis systems [1][2] gave rise to major improvements in Text-to-Speech (TTS) during the last 15 years. 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 creates 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 in a minimal set of sentences. It is traditionally the result of an optimization process, which raises the two following problems: 1- Which optimization algorithm should be used? 2- Which phonetic and prosodic criteria are best suited for this stage? In this paper we focus on the latter problem.

We report here a preliminary study concerning a novel phonetic unit, called a "vocalic sandwich," designed specifically for activities around unit selection TTS, like corpus design or perceptual quality prediction. Although it is no substitute for the basic concatenative units used in the TTS engine (diphones, half-phones…), vocalic sandwiches coverage is intended to provide better description abilities of TTS quality than traditional units. This study can be seen as a generalization of the work presented in [3].

In the following section we define the vocalic sandwiches and describe their general features. In section 3 we present an objective comparison of this unit with more traditional units. In section 4 we give a perceptual evaluation based on correlations between units coverage rates and Mean Opinion Scores. In the last section we summarize our contributions and present future work.

2. Vocalic sandwiches

2.1. Motivation

Diphones are often considered as the most appropriate units for concatenative TTS. Indeed, concatenating on stable areas of speech minimizes distorsion and preserves natural transitions between phones. Consequently, most unit selection TTS engines use diphones as their base units. In the following study we exclusively use the Orange Labs (ex- France-Télécom R&D) diphone-based system.

Nonetheless diphones prove to be inappropriate for corpora design. Since corpus-based TTS makes use of original speech segments of typically 3-4 phonemes, it is common to take longer units into account when preparing the script: triphones, words, etc. We did not find any dedicated analysis of such units in the specific area of corpus-based TTS. In other words, the impact of these approaches, in terms of final speech quality, has to be explored.

When designing textual corpora, concatenative TTS own features have to be taken into consideration. More specifically, such systems frequently suffer from acoustic glitches occurring around joins. The spectrum, energy or pitch mismatch responsible for such artefacts 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 [4]
- *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

To take account of these factors, several refinements of the concatenation cost used in the selection process were introduced [5][6]. A simpler -but still very efficient- criterion consists in penalizing concatenations depending on the phoneme type on which they occur (for instance vowel < liquid < semi-vowel < consonant) [7]. We suggest hereafter to adapt this "runtime" principle in a new phonetic unit, which may prove useful especially in the corpus design stage.

2.2. Definition

We use the term "Vocalic sandwich" to denote every phoneme sequence matching the following expression:

$$S \quad \overline{S}^{+} \quad S \qquad (1)$$

where :

- S denotes a set of phonemes where splicing is generally acceptable
- $\overline{S}$ refers to the complementary of S, *i.e.* a set of phonemes considered as inappropriate for splicing
- + is the regular expression quantifier meaning "one or more times"

The content of S may be adapted to the purpose. It typically contains most consonants (stops, fricatives, nasals) as well as the silence phoneme. In the experiment presented in section 4 we also include liquids [l] and [ʁ][1] in S, although they can cause uncertain splicing quality, since their voicing characteristic may be affected by surrounding phonemes. For very high quality TTS and large corpora, it is preferable to exclude them from S.

The term *vocalic sandwich* represents two "robust" phonemes surrounding a sequence of "fragile" phonemes like vowels and semi-vowels, which are consequently preserved from concatenations. As most co-articulation effects rarely cross consonants, they are also contained in sandwiches. As a consequence, these units are expected to be appropriate for many applications using concatenative speech synthesis.

We found early signs of such approach in [8], where "syllable-like units" derived from speech recognition [9] were favorably used for Indian Tamil TTS. They were delimited by minimum energy regions detected with a group-delay based algorithm. These acoustic units were then used to constrain the selection process, as could probably be done in a diphone system using an adequate concatenation cost. However such highly signal-dependent units cannot meet more general requirements like corpus design, where phonetic or linguistic considerations are necessary.

Definition (1) covers two types of units, both of which will not, however, be treated on equal terms. In fact, only units containing a vowel will be considered as actual "vocalic sandwiches." The others are commonly called "consonant clusters." Moreover, it follows from our diphone approach that boundary phonemes (set S) belong to two consecutive sandwiches or one sandwich and one cluster. Consequently, every phonetic sequence delimited by two silence phonemes can be split in a series of vocalic sandwiches and consonant clusters, as illustrated in table 1. This example also shows that sandwiches can extend across word boundaries.

(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)	#es kɛn ʁaɛk sjɔn səwik səʁ sɛp nɛl
(4)	nds ks ps 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 our work, almost no attention is paid to consonant clusters since they support concatenations very well, and their coverage is not really an issue (around 200 distinct clusters in French).

In section 4 we will also use sandwich 2-grams, consisting of two consecutive sandwiches possibly separated by a consonant cluster. The example in table 1 is associated to the following 2-gram list: (null - #es), (#es – səwik), (səwik - kɛn), (kɛn - nds - səʁ), (səʁ - ʁaɛk), (ʁaɛk - ks - sɛp), (sɛp – ps - sjɔn), (sjɔn - nɛl), (nɛl - l# - null)

2.3. Context-dependent sandwiches

In many applications, the phonetic content alone may be insufficient to characterize a sandwich. We suggest using a context-dependent version of these units, including information about the linguistic and prosodic context. To avoid dispersion of the overall sandwich distribution, we retain contextual data only for the main prosody carriers, namely the vowels. Furthermore we reduced the set of possible contexts to 13, which seems to be sufficient for French neutral speech.

3. Objective comparison

3.1. Unit statistics

In table 2 we give statistical data concerning vocalic sandwiches and traditional units. Measurements were done on a French textual corpus of 179,000 words in the domain of Interactive Voice Response (IVR) applications. Like sandwiches, traditional units were also derived in a context-dependent version (see 2.3). For each type of unit, measured values are:

- *Mean length*: the average number of phonemes included in one unit
- *Distinct units*: the total number N of distinct units encountered in the corpus
- *80% coverage*: the number of distinct most frequently encountered units needed to cover 80% of the corpus
- *Density* (in units/phoneme): ratio between the number of units and the number of phonemes in the full corpus
- *Overlap ratio*: measures how much two consecutive units "phonetically" overlap, on average, in proportion to the first unit
- *Entropy* (in bits/unit): quantifies the dissemination of units. Entropy E is maximal for uniform distributions, and increases with the number of distinct units:

$$E = -\sum_{i=1}^{N} f_i \cdot \log_2 f_i \qquad (2)$$

where f_i denotes the frequency of the i-th unit.

Depending on the way silences are treated, some of these indicators may differ slightly from expected values.

Since all unit distributions are governed by the Zipf-Mandelbrot law [10][11], the *distinct units*, *80%-coverage* and *entropy* columns give us valuable indications on the relative dissemination of the different units. A unit that would be easy to cover should show a "compact" distribution, referring to low values for these indicators. Logically, these indicators increase with unit complexity, which is related to the unit length and the amount of contextual information.

A "good" unit is therefore a compromise between compactness and complexity, the latter being associated to long-term and accurate description abilities, which should be related to perceived speech quality (see section 4 for perceptual evaluation).

[1] IPA notation (International Phonetic Alphabet)

	unit	mean length	distinct units	80% coverage	density	overlap ratio	entropy
context-dependent	phone 1-grams	1	214	43	1,00	0,00	13,75
	phone 2-grams	2	6 677	1 001	1,00	0,50	24,25
	phone 3-grams	3	52 570	7 724	0,95	0,65	30,66
	phone 4-grams	4	135 483	30 801	0,91	0,73	34,51
	phone 5-grams	5	215 956	74 651	0,86	0,77	36,91
	phone 6-grams	6	277 822	135 532	0,82	0,80	38,50
	phone 7-grams	7	321 187	186 504	0,78	0,82	39,60
	phone 8-grams	8	349 599	222 351	0,73	0,83	40,39
	syllable 1-grams	2,4	9 255	876	0,43	0,00	23,65
	syllable 2-grams	4,7	68 581	18 361	0,38	0,44	32,62
	word 1-grams	4,6	28 966	4 878	0,22	0,00	27,07
	word 2-grams	9,0	74 838	46 903	0,17	0,34	35,28
	sandwich 1-grams	3,2	21 690	2 834	0,38	0,18	27,54
	sandwich 2-grams	5,3	102 167	32 582	0,43	0,56	33,98
context-independent	phone 1-grams	1	35	18	1,00	0,00	11,00
	phone 2-grams	2	1 116	252	1,00	0,50	19,96
	phone 3-grams	3	14 494	2 139	0,95	0,65	26,87
	phone 4-grams	4	66 991	11 221	0,91	0,73	31,76
	phone 5-grams	5	147 911	37 410	0,86	0,77	34,95
	phone 6-grams	6	220 487	79 606	0,82	0,80	37,01
	phone 7-grams	7	273 680	138 997	0,78	0,82	38,47
	phone 8-grams	8	309 988	182 740	0,73	0,83	39,50
	syllable 1-grams	2,4	3 135	220	0,43	0,00	19,30
	syllable 2-grams	4,7	42 041	7 371	0,38	0,44	30,13
	word 1-grams	4,6	18 516	2 028	0,22	0,00	24,78
	word 2-grams	9,0	63 959	36 024	0,17	0,34	34,12
	sandwich 1-grams	3,2	10 270	923	0,38	0,18	24,22
	sandwich 2-grams	5,3	78 269	18 699	0,43	0,56	32,39

Table 2. Statistics of Vocalic Sandwiches and some traditional units.

3.2. Discussion

Table 2 indicates that sandwich 1-grams and 2-grams have an average length of 3.2 and 5.3 phonemes, respectively. According to their *80%-coverage* and *entropy* values, their distribution appears to be quite compact when compared to phone n-grams of similar length. This is an interesting feature in corpus design: sandwiches should be appropriate for small corpora while taking into account relatively long-term phenomena. Another feature that sets sandwiches apart from phone n-grams is their low *overlap ratio*, which offers flexibility for the selection or creation of uncovered unit sequences in order to create dense corpora.

We can also see from table 2 that the context-dependent versions of the units increase dispersion compared to the context-independent versions, approximately by a factor of 3 for the *80%-coverage* measure. The restriction to 13 contexts, carried only by vowels, probably minimized this increase.

4. Perceptual evaluation

The objective comparison of preceding section indicates that sandwiches demonstrate some interesting properties to be used in corpus design. Since their definition relies on perceptual observations, we also expect them to be well correlated to perceived speech quality. Supposing that their suitability for the corpus design stage is strongly related to their ability to predict TTS quality given the input sentence, we conducted the following perceptual evaluation.

4.1. Data collection

Evaluation data was collected within the CLASSiC European project, using the state-of-the-art diphone unit selection Orange Labs speech synthesizer. The speech synthesizer uses a female French voice tailored to the IVR domain: its acoustic inventory is mainly composed of utterances in the domain of IVR applications. During the recordings, the professional speaker was not constrained to speak in any neutral style and the voice is thus more expressive than most of the usual TTS voices, with a relatively high F_0 standard deviation (62 Hz, 4.6 semi-tones). Due to its relatively limited domain and its expressive intonation, the quality of the synthesized utterances is highly influenced by the closeness of the utterance to the acoustic inventory.

144 different utterances were taken in equal proportions from two application domains: IVR applications (e.g. “J’annule votre demande.”- I cancel your order.), and movie subtitles (e.g. “Tu me donnes quel âge ?”- How old do you think I am?).

Furthermore, each of the utterances was synthesized with two different versions of the TTS voice: a version that uses the full acoustic inventory (approximately 3 hours of speech) and one that uses a reduced acoustic inventory (30% of the full inventory, approximately 1 hour of speech, selected randomly). In total, the 288 synthesized utterances show a wide range in speech quality.

Twelve French native naïve listeners were each asked to rate the speech quality of 48 synthesized utterances on a MOS (Mean Opinion Score) 1 (bad) to 5 (excellent) scale. Each of the 288 synthesized utterances was so rated by two listeners and was given the average of the two individual scores.

4.2. Feature extraction

Coverage rates were extracted for each synthesized utterance and unit type. The coverage rate $C(u_1,...,u_N, A)$ of a sequence of N units $u_1,...,u_N$ for an acoustic inventory A is given by:

$$C(u_1,...,u_N, A) = \frac{1}{N}\sum_{i=1}^{N} \delta(u_i, A) \quad (2)$$

where $\delta(u_i, A)$ equals 1 if u_i is present in A and 0 if u_i is not present in A. For a sequence, the coverage rate ranges from 0 when no unit is covered in the inventory to 1 when all the units are present in the inventory.

At the same time, the unit selection cost of the TTS engine was also extracted for each synthesized utterance.

4.3. Results

We measured the correlations between the MOS and the coverage rates of each unit type on the whole data set. The correlation coefficients are reported on Figure 1 with 95% confidence intervals for the 14 unit types, context-dependent and context-independent. For information, MOS of both TTS voices on both application domains are also given in table 3.

As expected, all the correlation coefficients are positive, since high coverage rate is related to high MOS. The maximum correlation is 0.57, obtained for the context-independent sandwich 2-grams. When only comparing the context-dependent units, sandwich 2-grams are also ranked first, with a correlation of 0.53. However, given the significance intervals, this ranking should be considered with caution. Other units, like phone 5- and 6-grams show close correlations; this is in accordance with the average length of sandwich 2-grams (5.3 phonemes). It is interesting to note that this "optimal" length is longer than the length of most units traditionally used for corpus design, typically phone 3-grams.

For comparison, the correlation coefficient between the MOS and the unit selection cost is -0.48, which is of course highly dependent on the cost function used in the Viterbi algorithm. In terms of absolute value, it is significantly lower than the correlation with sandwich 2-grams. This is most likely due to the short-term limitations of the cost function, although it benefits from the acoustic features of the units.

Note that we have also extracted other features for the different unit types, like language model perplexities or number of un/covered units. From all the features, the coverage rates gave the best correlations.

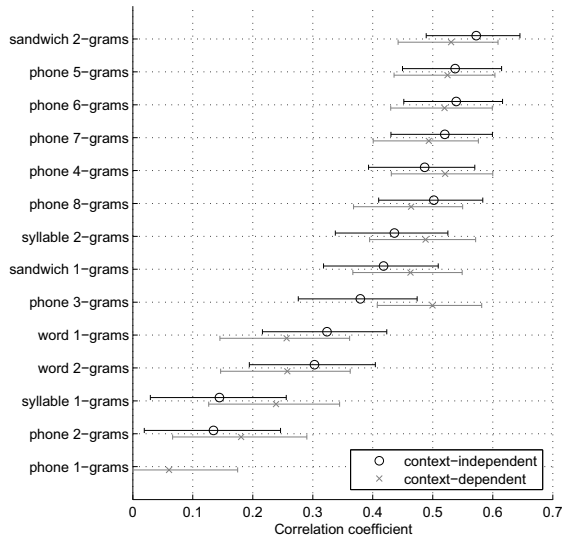

Figure 1: *Correlation coefficients between MOS and coverage rates of the different unit types.*

	IVR	Movie subtitles
Full voice (3h)	4.0	3.2
30% voice (1h)	3.5	2.9

Table 3. *Some mean opinion scores.*

4.4. Discussion

Table 4 reports mean coverage rates as well as their standard deviation for all unit types.

	context-independent		context-dependent	
unit	mean	std	mean	std
sandwich 2-grams	0.51	0.23	0.41	0.23
phone 5-grams	0.48	0.25	0.35	0.26
phone 6-grams	0.34	0.26	0.25	0.24
phone 7-grams	0.25	0.24	0.18	0.22
phone 4-grams	0.72	0.19	0.51	0.24
phone 8-grams	0.18	0.22	0.13	0.20
syllable 2-grams	0.57	0.22	0.40	0.24
sandwich 1-grams	0.91	0.10	0.78	0.18
phone 3-grams	0.95	0.06	0.78	0.16
word 1-grams	0.73	0.18	0.49	0.18
word 2-grams	0.21	0.18	0.12	0.14
syllable 1-grams	0.93	0.08	0.86	0.11
phone 2-grams	1.00	0.00	0.97	0.05
phone 1-grams	1.00	0.00	1.00	0.01

Table 4. *Mean coverage rates and standard deviations of the different unit types.*

For a unit to be discriminating and have high perceptual correlation, it should preferably yield a medium value for its mean coverage rate, typically between 0.2 and 0.8, as well as a high standard deviation. For example, units like sandwich 1-grams or phone 3-grams obtain better correlations in their context-dependent form than in their context-independent form. Indeed their mean coverage rates decrease from 0.91/0.95 to 0.78 when adding contextual information, and their standard deviation increases consequently.

We also computed correlations on several subsets of the collected data: every application domain (IVR / Movie Subtitles / Both) was intersected with every acoustic inventory version (Full / Reduced / Both). In all cases, context-independent sandwich 2-grams rank first, except for the IVR-Full subset where their context-dependent version is ahead. This observation is consistent with the fact that coverage rates are maximal on this subset. Similarly, we noticed a progression of short and context-independent units on the MS-Reduced subset, where coverage rates are minimal.

The expressiveness of the speech database used for this experiment probably justifies the overall predominance of long units, since long-term prosodic coherence may be more critical than for neutral voices. It could also explain the relatively low performance of context-dependent units, which were expected to perform better than their context-independent counterpart; indeed, contexts were tailored for neutral speech and seem to be of limited help for such expressive voice. Consequently it would be relevant to carry out similar evaluation with a more neutral voice. Comparison of results may be very instructive.

5. Conclusions

We introduced in this paper a new phonetical unit, called a "vocalic sandwich", which takes account of specific features of diphone-based unit selection TTS. It demonstrates interesting properties and may prove useful especially in corpus design. However further investigation is needed to explore its potential. Among other things, it would be interesting to:

- reproduce the perceptual experiment on voices of various speaking styles,
- use vocalic sandwiches for actual applications like corpus design and database reduction,
- validate their relevance for other languages.

Preliminary experiments we conducted in corpus design gave very promising results and will be object of future work.

6. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 216594 (CLASSIC project: www.classic-project.org).

7. References

[1] Sagisaka Y., "Speech synthesis by rules using an optimal selection of non-uniform synthesis units", ICASSP'88, pp. 679-682

[2] Hunt A., Black A., "Unit selection in a concatenative speech synthesis system using a large speech database", ICASSP'96, pp. 373-376

[3] Cadic D., Segalen L., "Paralinguistic elements in speech synthesis", Interspeech 2008

[4] Lindblom B., "Spectrographic study of vowel reduction", JASA, 35, pp. 1773-1781, 1963

[5] Donovan R.E., "A new distance measure for costing spectral discontinuities in concatenative speech synthesizers", ESCA 2001

[6] Bulyko I., "Flexible speech synthesis using weighted finite state transducers", PhD Thesis, University of Washington, 2002

[7] Yi J., Glass J., "Natural-sounding speech synthesis using variable-length units", ICSLP 1998

[8] Thomas S., Rao M.N., Murthy H.A., Ramalingam C.S., "Natural sounding TTS based on syllable-like units", EUSIPCO 2006

[9] Hu Z., Schalkwyk J., Barnard E., Cole R., "Speech recognition using syllable-like units", ICSLP'96, pp. 1117-1120

[10] Van Santen J., "Combinatorial issues in text-to-speech synthesis", EUROSPEECH'97, pp. 2511-2514

[11] Zipf G. K., "Selective studies and the principle of relative frequency in language", Harvard University Press, 1932