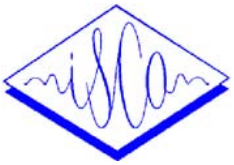

5th International Conference on Spoken Language Processing (ICSLP 98)
Sydney, Australia
November 30 - December 4, 1998

NATURAL-SOUNDING SPEECH SYNTHESIS USING VARIABLE-LENGTH UNITS[1]

Jon R.W. Yi and James R. Glass

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
{jonyi, jrg}@sls.lcs.mit.edu

ABSTRACT

The goal of this work was to develop a speech synthesis system which concatenates variable-length units to create natural-sounding speech. Our initial work in this area showed that by careful design of system responses to ensure consistent intonation contours, natural-sounding speech synthesis was achievable with word- and phrase-level concatenation. In order to extend the flexibility of this framework, we focused on the problem of generating novel words from a corpus of sub-word units. The design of the sub-word units was motivated by perceptual studies that investigated *where* speech could be spliced with minimal audible distortion and *what* contextual constraints were necessary to maintain in order to produce natural sounding speech. The sub-word corpus is searched during synthesis using a Viterbi search which selects a sequence of units based on how well they individually match the input specification and on how well they sound as an ensemble. This concatenative speech synthesis system, ENVOICE, has been used in a conversational information retrieval system in two application domains to convert meaning representations into speech waveforms.

10.21437/ICSLP.1998-575

1. INTRODUCTION

In an ideal world, a speech synthesizer should be able to synthesize any arbitrary word sequence with complete intelligibility and naturalness. The trade-off schematic in Figure 1 illustrates how current synthesizers have tended to strive for flexibility of vocabulary and sentences at the expense of naturalness (i.e., arbitrary words and sentences can be synthesized, but do not sound very natural.) This applies to articulatory, rule-based, and concatenative methods of speech synthesis [2, 5, 6, 9].

An alternative strategy is one which seeks to maintain naturalness by operating in a constrained domain. There are potentially many applications where this mode of operation is perfectly suitable. In conversational systems for example, the domain of operation is often quite limited, and is known ahead of time. An extreme example of obtaining naturalness is the use of pre-recorded speech. A step beyond this is word- or phrase-level concatenation of speech segments from pre-recorded utterances. As we wish to increase word flexibility, we turn to concatenating ever smaller-sized units. The decision of which units to use in concatenative synthesis is a process guided by contextual information to preserve co-articulatory and prosodic constraints.

[1]This research was supported by DARPA under Contract N66001-96-C-8256, monitored through Naval Command, Control and Ocean Surveillance Center, and by a research contract from BellSouth Intelliventures.

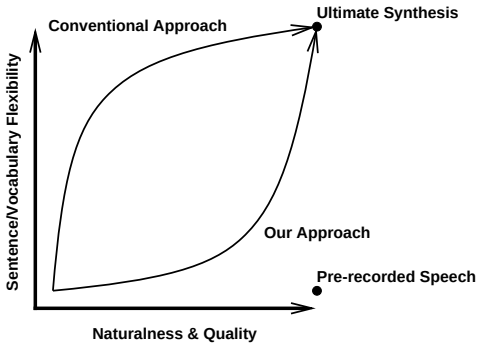

Figure 1: Synthesis development trade-off schematic.

Past works by others have examined how unit selection algorithms can be formulated, and what constraints must be maintained [2, 6, 9].

In this work, we develop a framework for natural-sounding speech synthesis using variable-length units. The developmental philosophy that we have adhered to throughout the work places naturalness as the paramount goal. We achieve this in constrained application domains by performing Meaning-to-Speech (MTS) synthesis directly, avoiding a potentially lossy intermediate text re-analysis step. In our preliminary work involving word- and phrase-level concatenation, the vocabulary size is relatively small, but naturalness is quite high. After the sub-word architecture is developed, new words can be manufactured from a sub-word unit database by naturally concatenating possibly non-contiguous speech segments. Our research follows the bottom curve in Figure 1 where we view naturalness as the highest priority, while steadily increasing sentence and vocabulary flexibility. As the pursuit of naturalness dominates, human listening provides the best feedback.

2. PHRASE-LEVEL CONCATENATION

This section describes preliminary work that led up to the development of a sub-word concatenation framework. The synthesis process involves the concatenation of word- and phrase-level units with no signal processing. These units are carefully prepared by recording them in the precise prosodic environment in which they will be used. Using carrier phrases as vehicles for recording words and phrases and as the basis for synthesis, this type of unit design and unit concatenation achieves a high level of naturalness.

In conversational systems, responses can be generated from an internal, high-level meaning representation or *semantic frame*. These frames store information as key-value pairs which can be possibly recursive, containing other frames. In this work, we make use of GENESIS [4], a language generation system that recursively builds a response given a meaning representation and a message template. This message template then draws on elements from the frame and run-time lookups from a pre-defined vocabulary to generate the sentence. We use this lookup ability to insert ENVOICE-specific annotations specifying waveform segments. In this response generation framework, a meaning representation is converted by GENESIS to an annotated description of waveform segments for concatenation by ENVOICE.

In a classified ad domain for used cars [8], phrase-level concatenative synthesis was used to describe automobile advertisements that a user would encounter during a typical dialog. For example, given the appropriate message templates and vocabulary items, the database record seen in Figure 2 can be synthesized into the three following sentences. Note that arbitrary number generation is possible using phrase-level units (e.g., year, mileage, price, telephone number.) [SOUND 1151_01.WAV] [SOUND 1151_02.WAV] [SOUND 1151_03.WAV]

```
{ Wheels
   :year 1996
   :color "black"
   :model "integra"
   :make "acura"
   :mileage {q number :1 40 :2 5000 :3 300 :4 and
     :5 80 }
   :price {q number :1 8000 :2 900 :3 and :4 70 }
   :telno {q telephone :1 "1_4" :2 "2_0" :3 "3_4"
     :4 "4_3" :5 "5_9" :6 "6_9" :7 "7_7" :8 "8_6"
     :9 "9_8" :0 "0_2" }
   :nth_ad "third" }
```

The third ad is a 1996 black Acura Integra with 45,380 miles.
The price is 8,970 dollars.
Please call (404)399-7682.

Figure 2: MTS example in a used-car domain.

By carefully designing the corpus for response generation, it is possible to achieve very natural-sounding concatenated speech in a constrained application domain. However, recording every word in every prosodic environment realizable represents a trade-off between large-scale recording and high naturalness. Essentially, this type of generation approach has two shortcomings. First, while the carrier phrases attempt to capture prosodic constraints, they do not explicitly capture co-articulatory constraints, which may be more important at sub-word levels. Second, some application domain vocabularies are continuously expanding (e.g., new car models may be introduced each year), or have a large number of words (e.g., the 23,000 United States city names in a Yellow Pages domain).

More generally, proper names are types of words that can potentially have large set sizes which may also grow as time goes on. While brute-force methods would dictate the recording of every word, we decided to investigate methods that concatenate sub-word units from a designed sub-word corpus for the synthesis of arbitrary proper names.

3. PERCEPTUAL STUDIES

In this section, we report on the results of some perceptual studies in which we attempted to learn what units are appropriate for concatenative synthesis, and how well these units sound as an ensemble when concatenated together to form new words. We call these two constraints the *unit* and *transition* criteria [6]. These tests not only describe how much contextual information is required, but also where the unit boundaries should lie.

Because source changes (e.g., voiced-unvoiced transitions) typically result in significant spectral changes, we hypothesized that a splice might not be as perceptible at this point, in comparison to other places. Should the speech signal be broken between two voiced regions, it would be important to ensure formant continuity at the splice boundary. This hypothesis motivated a series of consonant-vowel-consonant (CVC) studies that dealt with the substitution of vowels at boundaries of source change.

Transition criterion In the CVC studies that tested potential transition points, we fixed the place of articulation of the surrounding consonants. For example, the /ɔ/ from the city name, "Boston" (/b ɔ s t ɪ n/), was replaced by the /ɔ/ from "bossed" (/b ɔ s t/). Perceptually the splicing is not noticeable. We found this effect to hold when the consonants are stops, fricatives, as well as affricates.

As a variation on the previous study, we used the /ɔ/ from the word "paucity" (/p ɔ s ɪ t i^y/), changing the voicing dimension of the stop consonant on the left side of /ɔ/. Because part of the /ɔ/ from "paucity" is devoiced, there is less formant transition in the voiced part of the vowel. However, perceptual listening indicated that this was a secondary effect and that overall the splicing still sounded natural. This knowledge contributed towards the formation of the unit criterion studies.

Unit criterion In extending the above study to testing contextual constraints, our initial hypothesis was that place of articulation was an important context to maintain in sub-word unit concatenation. We created three classes to capture this triphone context: labial, alveolar/palatal/dental, and velar. We found alveolar (t d̪ s z), dental (θ ð), and palatal (š ž č ǰ) contexts to have similar effects on formant frequency locations and to usually produce natural-sounding synthesis.

In a study involving nasal contexts, we kept the place of articulation constant in order to isolate the phenomenon of nasalization. For example, when the /æ/ from "map" is substituted into "pap", the nasalized vowel does not sound natural. Nasalization of a vowel occurs when a consonant to either side is a nasal, and we will not be able to dismiss it from the unit criterion definition as we did with voicing. Continuity in vowel nasalization and nasal murmur across vowel-nasal boundaries is important to maintain [3]. We found this effect to be stronger for vowel-nasal sequences than for nasal-vowel sequences, possibly because anticipatory nasalization is a stronger effect in English.

In summary, we found the place of articulation and nasal consonants to be the main contextual constraints for vowels. While it was possible to perform natural-sounding splicing at boundaries between vowels and consonants, we found it preferable to keep vowel and semivowel sequences together as a unit.

4. DESIGN OF SYNTHESIS UNITS

In this section, the various principles learned from the perceptual studies are used to enumerate a set of synthesis units for concatenative synthesis of non-foreign English words. As part of the study, a set of words are automatically generated to serve as a sub-word corpus that compactly covers the co-articulatory inventory required for unconstrained synthesis. Since the space of units can grow large as context is added, the use of linguistic knowledge can help to reduce the number of contexts [1].

To determine the units required for synthesis, we made use of a 90,000-word lexicon from the Linguistic Data Consortium called the *COMLEX English Pronunciation Dictionary*, commonly referred to as PRONLEX. We limited our analysis of contiguous multi-phoneme sequences of vowels (V) and semivowels (S) to the non-foreign subset of PRONLEX containing approximately 68,000 words. Consonant sequences and lexical stress were ignored. The result of this analysis is shown in Table 1.

Length	Example	# of units	# of occurrences	% cumulative
1	ɪ	16	100513	63.3
2	rɪ	167	40778	89.0
3	ərɪ	484	12916	97.1
4	yəleʸ	637	3392	99.3
5	oʷrɪʸəl	312	906	99.8
6	yələrɑʸ	80	226	100.0
7	ærəlɛlɪ	13	21	100.0
Total		1709	158752	100.0

Table 1: PRONLEX analysis of vowel and semivowel sequences.

We prepared a unit inventory by first expanding the sequences with triphone consonant and silence contextual information. Next, we compressed the triphone context using seven contextual classes learned from the perceptual studies: labial, alveolar/dental, velar, m, n, ŋ, and silence. The numbers tabulated from these two stages are shown in Table 2 along with coverage statistics for the units from the final stage.

Sequence length	1	2	3	4	5	6-7
# of units (stage 1)	2970	4814	3883	1643	515	126
# of units (stage 2)	541	1817	2343	1342	463	113
% unit coverage	63.3	89.0	97.1	99.3	99.8	100.0
% word coverage	29.8	75.2	93.3	98.3	99.6	100.0

Table 2: Comparison of sequences and coverage.

Because most longer multi-phoneme sequences of vowels and semivowels occur in only a small number of the words within the lexicon, we chose an operating point of a sequence length of 2 (i.e., VV, SV, and VS sequences). To synthesize any non-foreign PRONLEX word, we need 2,358 unique vowel and semivowel sequences; consonants are assumed to be adequately covered. While these sequences could be covered using a brute-force approach by recording a word for each sequence, we used an automatic algorithm to select a compact set of prompts to record given a set of units to cover and a set of words to choose from.

The prompt selection algorithm selects the next best word to incrementally cover the most infrequent units remaining to be covered without providing redundant units [7]. Ties are broken randomly in this iterative process until all units have been covered. When this prompt selection algorithm was applied, a total of 1,604 words was selected.

5. UNIT SELECTION

The unit selection algorithm is a Viterbi search that provides an automatic means to select an optimal sequence of sub-word units from a speech database given an input pronunciation. Using speech knowledge as encoded by the researcher, the search metric seeks out units that individually match the input specification well and that connect well as an ensemble. These two criterion can be decoupled and separately considered as a unit cost function and a transition cost function. Because the use of longer-length units tends to improve synthesis quality [2, 9], it is important to maximize the size and the contiguity of speech segments to encourage the selection of multi-phone sequences. The speech database is only phonetically time-aligned, and, therefore, this late-binding approach will seek out desirable sub-word units when present and can back-off to shorter units when necessary. If supervised selection is a possibility, an A^* search can be used to obtain an N-best list of synthesis paths.

The unit cost function measures co-articulatory distance by considering triphone classes which have consistent manner of production: vowels/semivowels, stops, nasals, silence, and a final group that includes fricatives, affricates, and the aspirant, /h/. For vowels, place of articulation and nasal consonant contexts are important factors. Allophonic variations of stops are mainly attributed to flapping and the presence of front, back, round, or retroflexed environments [11]. For fricatives, we consider round and retroflexed environments, whereas for nasals, the constraints deal with syllable position (onset or coda) and contexts producing durational lengthening. As voiced consonants to the right of a nasal tend to give the nasal a longer duration [3], incorrect usage of allophonic variations of nasals can confuse the listener as to whether the following stop is voiced or not (e.g., a synthesized "bent" with a lengthened /n/ sounds like "bend.")

The transition cost function measures co-articulatory continuity between two phones proposed for concatenation. A transition cost must be incurred if they were not spoken in succession to avoid concatenations at places exhibiting a significant amount of co-articulation, or formant motion. To model higher-level constraints, we considered six manner classes: vowels, semivowels, nasals, /h/, obstruents, and silence. The reason for /h/ occupying a class by itself arises from some of our other perceptual studies showing that it adapts to its co-articulatory environment.

We decouple transitions occurring within or across syllables into intra-syllable and inter-syllable transitions, respectively. The cost matrices implicitly encode many types of knowledge including speech production and sub-syllabic structure. For example, VV, SV, and VS sequences can be preserved with high transition costs. A high obstruent-obstruent intra-syllable transition cost helps to capture allophonic variations of obstruents in clusters such as unvoiced stops which are unaspirated in onset and coda clusters with /s/ (e.g., stop, spots.) In another example involving retroflexion, /stɹ/, this same transition cost will encourage the selection of a retroflexed /s/ adjacent to a /t/ selected from a retroflexed environment. Thus, longer-distance constraints can even be captured with just transition and triphone unit costs. As a final note, the inter-syllable costs are generally lower than intra-syllable costs, because we observed that contiguity preservation is more important within a syllable than across syllables.

6. SYNTHESIS EXPERIMENTS

With a concatenative synthesis framework incorporating sub-word and phrase-level units, we conducted experiments involving isolated words and full sentences. Using sub-word synthesis we generated novel proper names using sub-word units from a corpus of common words. In another example, we combined sub-word and phrase-level synthesis to synthesize sentences in which we back off to sub-word synthesis for novel words.

Sub-word synthesis was used to synthesize city names from a corpus of common, non-foreign English words. Because city names often have foreign etymologies, a database of common words may not provide enough co-articulatory richness. (However, we also note that some of the common words of a given language may have originally been imported from elsewhere.)

The testing data set was a list of 485 cities from a weather information domain [10]. The training data set of 318 common words was formed by running the selection algorithm using the non-foreign subset of PRONLEX to cover the 485 city names. It was recorded by a native American-English female speaker and phonetically time-aligned using a speech recognizer. Then, the phonetic labels were collapsed into phonemic labels (e.g., stop closure and release collapsed into stop) to better match PRONLEX pronunciations. Another set of transcriptions were prepared containing syllable boundaries that were automatically determined using a simple rule-based syllabification algorithm we designed.

In Figure 3, we present one of the 485 city names, "Acapulco", where square brackets are used to denote which sub-word portions were selected. The syllabification of "Acapulco" is: (æ) (kə) (pʊl) (kʊʷ). [SOUND 1151_04.WAV]

Acapulco				
[a]cclamations	tele[co]nnect	[p]oorhouse	f[ul]crum	pe[koe]
/æ/	/kə/	/p/	/ʊl/	/kʊ/

Figure 3: "Acapulco": sub-word synthesis path

The final experiment we present here demonstrates the integration of phrase-level and sub-word concatenation. This example response comes from a displayless flight status information retrieval system we are developing. Square brackets are used to denote sub-word boundaries, and curly braces are used to denote phrase boundaries. [SOUND 1151_05.WAV]

{Continental} {flight} {46}{9}{5} {from} [G][reen][sb][oro] {is expected in} [Hali][f][ax] {at} {10}:{08}{pm} {local time}.

The total variable-length concatenative synthesis framework operates in a networked conversational system, where ENVOICE servers return speech waveforms to clients presenting meaning representations as input. Overall, users thought the system sounded natural and found sentences to be much preferable over those generated by *DECTalk*.

7. CONCLUSIONS AND FUTURE WORK

This work has three types of contributions: a framework for MTS concatenative synthesis, principles about sub-word unit design for concatenative synthesis, and sub-word corpus design.

This MTS framework is suitable for use in a conversational system because it was designed from the ground up for understanding domains as opposed to general-purpose Text-to-Speech synthesizers. Concatenating at phone boundaries seems to be more natural than our past experience with diphone synthesis. Designing and performing perceptual studies helped to further our understanding of the perceptual effects of concatenation.

There remains much future work in many areas including unit design, prosody, evaluation methods, and development strategies. Contextual constraints for consonant selection should be investigated. The naturalness of poly-syllabic words could be improved by incorporating stress into the unit design and prosody into the search metric as well into a post-processing step of prosody modification. An intonation contour could either be automatically generated using statistical means, for example, or be obtained from utterances spoken by a human. For regressively comparing perturbations in search metric weights, it is necessary to devise an objective evaluation measure. Finally, a semi-automatic framework is desired for rapid prototyping of synthesizers for new vocabulary, domains, and languages.

8. ACKNOWLEDGMENTS

We would like thank the following people who donated their voices to this work: Michelle Spina, Vicky Palay, and Tom Lee.

9. REFERENCES

1. E. C. Albano and P. A. Aquino, "Linguistic criteria for building and recording units for concatenative speech synthesis in brazilian portoguese," in *Proc. Eurospeech*, Rhodes, Greece, pp. 725–728, Sept. 1997.

2. N. Campbell, "CHATR: A high-definition speech re-sequencing system," *Acoustical Society of America and Acoustical Society of Japan, Third Joint Meeting*, Dec. 1996.

3. J. Glass, *Nasal Consonants and Nasalized Vowels: An Acoustic Study and Recognition Experiment*. S.M. thesis, MIT, Cambridge, MA, 1984.

4. J. Glass, J. Polifroni, and S. Seneff, "Multilingual language generation across multiple domains," in *Proc. ICSLP*, Yokohama, Japan, pp. 983–986, Sept. 1994.

5. X. Huang, A. Acero, J. Adcock, H. Hon, J. Goldsmith, J. Liu, and M. Plumpe, "Whistler: A trainable text-to-speech system," in *Proc. ICSLP*, Philadelphia, PA, pp. 2387–2390, Oct. 1996.

6. A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. ICASSP*, Atlanta, GA, pp. 373–376, May 1996.

7. R. Kassel, "Automating the design of compact linguistic corporation," in *Proc. ICSLP*, Yokohama, Japan, pp. 1827–1830, Sept. 1994.

8. H. Meng, S. Busayapongchai, J. Glass, D. Goddeau, L. Hetherington, E. Hurley, C. Pao, J. Polifroni, S. Seneff, and V. Zue, "WHEELS: A conversational system in the automobile classifieds domain," in *Proc. ICSLP*, Philadelphia, PA, pp. 542–545, Oct. 1996.

9. Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *Proc. ICASSP*, New York, NY, pp. 679–682, Apr. 1988.

10. V. Zue, S. Seneff, J. Glass, L. Hetherington, E. Hurley, H. Meng, C. Pao, J. Polifroni, R. Schloming, and P. Schmid, "From interface to content: Translingual access and delivery of on-line information," in *Proc. Eurospeech*, Rhodes, Greece, pp. 2227–2230, Sept. 1997.

11. V. W. Zue, *Acoustic Characteristics of Stop Consonants: A Controlled Study*. Sc.D. thesis, MIT, Cambridge, MA, 1976.