Automatic characterisation of quasi-syllabic units for speech synthesis based on acoustic parameter trajectories: a proposal and first results

Parham Mokhtari and Nick Campbell
JST-CREST at ATR-HIS Laboratories, Keihanna Science City, Japan
parham@atr.co.jp nick@atr.co.jp

1 Introduction

Compared with rule-based methods of speech synthesis, concatenative methods have in the last decade or so become more popular and commercially successful. This can be attributed mainly to the greater perceived naturalness afforded by using segments of pre-recorded utterances, thus preserving both the speaker's voice characteristics and other acoustic details which have so far eluded formalisation in a synthesis-by-rule framework.

On the other hand, concatenative systems generally lack the flexibility of expanding to either new voices or a wider range of speaking styles. In our JST-CREST project, we are collecting vast amounts of natural speech data recorded in everyday situations, with the aim of capturing as much of the expressive and speaking style variations as are likely to be used by a given speaker on a day-to-day basis. However, while speech recognition technology can be used to force-align phonetic transcriptions onto acoustic data, creation of a unit-database appropriate for synthesis is still labour-intensive and costly, requiring extensive manual checking of the segmentation and labelling.

To summarise, although concatenative synthesis affords a high degree of perceived naturalness, so long as the definition of the basic segments or units of speech is overly reliant on conventional phonetic labels, creation of a new unit-database or expansion of an existing one remains an arduous task.

2 Acoustic specification of quasi-syllabic units

2.1 Rationale

The preceding discussion motivates the need to transcend, or at least to not be bound by, the written phonetic transcription. In that vein, Öhman's (2000) commentary on the overbearing influence of grammatical-linguistic concepts in speech processing, while controversial and provocative, is an inspiring point of departure. One application of his advocacy to represent speech directly in acoustic terms rather than the traditional alphabetic or phonetic symbols, is indeed in the specification of the basic units to be used in speech synthesis: a more natural and more direct path to unit characterisation, selection, and concatenation might be found by relaxing the dependence on discrete, well-demarcated phonemes, and embracing a more continuous, acoustically-driven approach.

The first issue to address concerns the type of basic segment. While the *syllable* has long been regarded as one of the most fundamental units of spoken language, from a practical point of view it is problematic as there is still no consensus on a definitive specification at either the linguistic or acoustic level. By contrast, simple yet powerful methods of using prosodic parameters such as sonorant energy to locate *quasi*-syllables in the continuous speech stream, have been known for many years (e.g., Mermelstein, 1975). Although such automatic methods cannot be said to yield a syllabification in a strict phonetic-linguistic sense, they have the advantages of *automatic reproducibility* and *acoustic consistency*, both of which are important in defining units for synthesis.

The second issue concerns segment characterisation – a question that calls on over half a century of research on the "information-bearing elements of speech" (Peterson, 1952). Clearly, the cepstrum has proved the most popular acoustic parameter over the last two decades, owing mainly to its measurement robustness and superiority in automatic speech recognition. However, motivated by the need to succinctly capture both the acoustic-phonetic and -prosodic dynamics of quasi-syllabic, hence variable-length speech segments, it may not be entirely unreasonable to call upon the *low dimensionality* and the *articulatory and perceptual relevance* of the formants {F1, F2, F3, F4}; the energy within sonorant- and higher-frequency bands {SE, HFE}; and the fundamental frequency of voicing {F0}. In the next two subsections we describe our speech data and our methods of parameterising the dynamics of automatically-located quasi-syllables.

2.2 Speech data

In anticipation of the much larger amounts of more natural recordings now in preparation, our phonetically-labelled speech data comprise three stories read by an adult, female, native speaker of Japanese (Iida et al., 1998). The stories were designed to naturally evoke the emotions Anger, Joy and Sadness. Altogether, there are 1370 sentence-length utterances stored in separate files for independent processing.

2.3 Methods

Before segmenting an utterance, the following measurements are made: (i) contours of SE (within 60 Hz to 3 kHz for this speaker) and of HFE (within 3.4 kHz to 6 kHz) are measured (in dB) directly from FFT spectra; (ii) the F0 contour is measured (Hermes, 1988) then converted to the perceptual ERBR scale (Moore & Glasberg, 1983); (iii) the first four formant frequencies are estimated at each analysis frame, by linear transformations of the linear-prediction (LP) cepstrum (Broad & Clermont, 1989). Although formants are properly defined only in voiced speech, and the cepstrum-to-formant mapping is trained with a balanced set of vowel steady-states, the mapping is nevertheless applied to every frame, thereby yielding continuous, *quasi*-formant contours across each utterance (properties of which will be reported elsewhere). Similarly, a continuous F0-contour is obtained per utterance, simply by linear interpolation through unvoiced regions.

Next, the acoustic speech stream is quasi-syllabified using the convex-hull algorithm (Mermelstein, 1975) applied to the SE contour. As a result, each quasi-syllable is delimited by a significant local-minimum in sonorant energy; as seen later, this property alone helps considerably to minimise acoustic discontinuities at segment boundaries. The dynamics of each quasi-syllable are then characterised by computing the mean and the first few Fourier cosine-series coefficients of the contours of SE, HFE, interpolated-F0, and quasi-F1, -F2, -F3, -F4 within each segment; a greater number of coefficients helps increase the captured resolution of each contour.

3 Application to speech synthesis

Clearly, for full text-to-speech synthesis we will require a mapping from orthographic text input to the sequence of quasi-syllabic acoustic contour parameters. Such a mapping could be approximated by any of the rule-based methods proposed during the last 30-40 years (e.g., Holmes et al., 1964; Rabiner, 1968; Elovitz et al., 1976) and used, e.g., as a front-end to the well-known Klatt synthesiser. The sonorant energy might be approximated via a combination of frame-wise values of Klatt-synthesis parameters, then subjected to the convex-hull algorithm to yield a quasi-syllabification analogous to that performed on natural speech as described earlier. After appropriate scaling of F0 and the formants to approximate overall characteristics of our recorded speaker, each quasi-syllabic parameter contour would then be characterised by cosine-series coefficients, the unit-database searched for the closest match to each segment, and those units concatenated to yield synthesis based on natural speech.

Another possible application is speech-to-speech synthesis,

音声合成における音響的特徴の軌道に基づいた準音節
単位の自動パラメータ化の検討
○ パーハム モクタリ、 ニック キャンベル
JST/CREST ESP Project, ATR-HIS Labs, Kyoto.

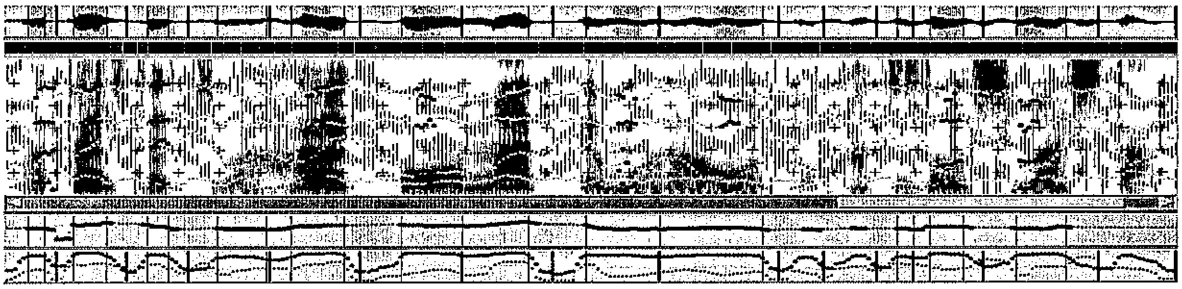

Figure 1. Waveform, spectrogram, F0- and energy-contours of an ***original utterance*** automatically quasi-syllabified (thick black vertical segmentation lines; thin vertical lines delimit quasi-syllabic nuclei). The four traces on spectrogram show quasi-F1, -F2, -F3, and -F4 estimated from LP-cepstra.

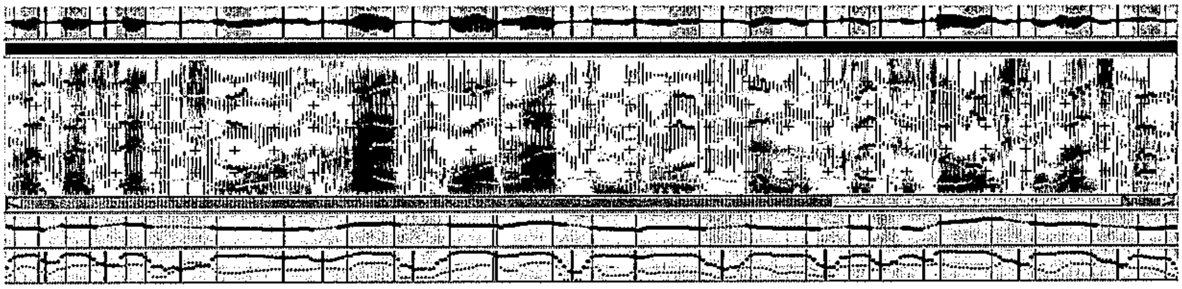

Figure 2. Waveform, spectrogram, F0- and energy-contours of an ***utterance synthesised by concatenating 15 segments*** which were found to best match the quasi-syllables in the original utterance in Figure 1. Graphical display shows results of post-concatenation re-analysis of the synthesised waveform.

where an input utterance from a different speaker is quasi-syllabified, parameterised, and transformed to the specifications required to generate the same utterance by concatenating quasi-syllables from the unit-database.

However, as a first, unbiased test of the proposed system, we use natural speech from the same, adult-female speaker, thus avoiding potential errors of acoustic parameter generation from text and subsequent speaker normalisation. In particular, we take as input only one of our speaker's utterances in turn, and search the remaining 1369 utterances for matching quasi-syllables which, upon concatenation, should ideally recreate the original utterance. Our present computation of the target-matching cost uses an inverse-variance-weighted Euclidean distance which therefore treats each parameter (including the quasi-syllabic duration) equally.

4 Preliminary results

Automatic segmentation of our database yielded 42,849 quasi-syllables, with a mean duration of 231msec. Ignoring segments either labelled as a silence (i.e., a pause) or with an unknown phonetic label, the remaining 34,154 units have a mean duration of 238msec, for a total of about 135 minutes of spoken data. However, even inter-speech pauses are included in our unit-selection in order to increase naturalness, as in a number of cases we observed that a matching *intake of breath* was correctly selected during a labelled silence!

Concerning the resolution at which the parameter-contour dynamics are captured, we found no significant gains beyond the first 3 or 4 Fourier cosine-series coefficients. This result is predictable, considering the general rate of acoustic change within an average quasi-syllabic duration of just over 0.2 sec. Amongst the formants, the F2 contours yielded about 3 times larger rms-errors of fit compared with F1, F3 and F4, as might be expected from the larger phonetic variations in F2.

Figures 1 and 2 compare an utterance and its resynthesis. As shown in Fig. 1, the original phrase "watashitachi wa haikoo o matte moraoo to katsudoo shimashita" was quasi-syllabified as follows: wa – tashI – ta – chiwa – hai – kooo – matt – ttemo – raoo – to – ka – tsu – doosh – shimashI – ta. In this example, the SE, HFE, F1 and F2 parameters were given ten times extra weight in unit-selection relative to those of F0, F3, F4, and duration. As seen in Fig. 2, the phonetic identities of 13 of the 15 segments were well reproduced: exceptions are the original segments "raoo" and "to", for which "roo" and "tano" were respectively selected. However, despite the per-unit phonetic accuracy, de-weighting of F0 and duration in the target-cost resulted in a prosodically choppy rendition; inverting the unit-selection weights generally yielded natural prosody at the expense of greater phonetic inaccuracies.

5 Discussion and ongoing research

Due to our encouraging preliminary results, we would like to improve on at least the following two points. First, there is the question of whether the acoustic parameterisation is sufficiently powerful to encode the phonetic and prosodic patterns of such quasi-syllables. As acoustic unit-selection admittedly resembles recognition, the cepstrum may help to select phonetically better-matched units (e.g., Zahorian & Jagharghi, 1993). Second, it may be assumed that relative to shorter, phone-based units, our quasi-syllabic inventory will imply much more data for comparable phonetic and prosodic coverage; on the other hand, our motivations for using such a unit include its acoustic consistency and measurement robustness in huge databases which also carry a natural range of paralinguistic variation. In ongoing research, we are exploring the combinatorics of phonetic/prosodic coverage, particularly in regard to speaking-styles; this task is greatly facilitated by our unsupervised methodology.

Acknowledgements

The first author would like to thank all members of the JST-CREST ESP project, in particular Carlos Toshinori Ishi and Kazuyuki Ashimura, for their encouragements and discussions.

References

Broad, D.J. & Clermont, F. (1989). "Formant estimation by linear transformation of the LPC cepstrum", *J.Acoust.Soc.Am* 86 (5), 2013-2017.

Elovitz, H.S., Johnson, R., McHugh, A. & Shore, J.E. (1976). "Letter-to-sound rules for automatic translation of English text to phonetics", *IEEE Trans. ASSP*-24, 446-459.

Hermes, D. (1988). "Measurement of pitch by subharmonic summation", *J. Acoust. Soc. Am.* 83 (1), 257-264.

Holmes, J.N., Mattingly, I.G & Shearme, J.N. (1964). "Speech synthesis by rule", *Language and Speech* 7, 127-143.

Iida, A., Campbell, N., Iga, S., Higuchi, F. & Yasumura, M. (1998). "Acoustic nature and perceptual testing of corpora of emotional speech", in *Proc. 5th Int. Conf. on Spoken Lang. Process.*, 1559-1562.

Mermelstein, P. (1975). "Automatic segmentation of speech into syllabic units", *J. Acoust. Soc. Am.* 58 (4), 880-883.

Moore, B.C.J. & Glasberg, B.R. (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", *J. Acoust. Soc. Am.* 74 (3), 750-753.

Öhman, S. E. G (2000). "Oral culture in the 21st century: the case of speech processing", in *Proc. Int. Conf. on Spoken Lang. Process.*, Beijing, China, 36-41.

Peterson, G.E. (1952). "The information-bearing elements of speech", *J. Acoust. Soc. Am.* 24, 629-637.

Rabiner, L. (1968). "Speech synthesis by rule: An acoustic domain approach", *Bell System Tech. J.* 47, 17-37.

Zahorian, S.A. & Jagharghi, A.J. (1993). "Spectral-shpe features versus formants as acoustic correlates for vowels", *J. Acoust. Soc. Am.* 94 (4), 1966-1982.