Quasi-syllabic and quasi-articulatory-gestural units for concatenative speech synthesis

Parham Mokhtari and Nick Campbell
JST-CREST at ATR-HIS Labs, Keihanna Science City, Kyoto, Japan
E-mail: parham@atr.co.jp, nick@atr.co.jp

ABSTRACT

In this paper we propose methods of speech segmentation and unit characterization which are motivated by prosodic and physiological principles. In particular, we motivate and describe algorithms for unit-database creation on the basis of *quasi-syllables* and *quasi-articulatory-gestures* defined and parameterized purely by acoustic measurements. This approach is intended to overcome the burden of reliance on the phonetic code in concatenative speech synthesis.

1. INTRODUCTION

In recent years, methods of speech synthesis based on the concatenation of appropriately selected segments (or units) have gained popularity over the more knowledge-based or rule-driven methods. This popularity is due mainly to the greater perceived naturalness afforded by the fact that the segments are taken from pre-recorded utterances and therefore preserve articulatorily- and perceptually-salient phenomena such as coarticulation, speaker individuality, and other acoustic details which have so far eluded formalisation in a synthesis-by-rule framework. However, despite their success, concatenative synthesizers have a number of serious drawbacks which have curtailed their practical use mainly to limited-domain applications.

One of the limitations of concatenative systems is that they generally lack the flexibility of expanding to either new voices or a wider range of speaking styles. This lack of flexibility can partly be attributed to the over-reliance on the phonemic transcription of spoken utterances, as part of the process of delimiting the units. Indeed, one of the major costs of preparing a unit-database is in the labour-intensive tasks of phonetic segmentation and labeling. Automatic speech recognition or forced-alignment methods are often used, but even after adaptation to the given speaker and speaking style the segmentation results must usually be carefully checked and errors corrected manually to ensure sufficiently high-quality synthesis. Given the huge amounts of speech data required to construct a unit-database with sufficient phonetic coverage (let alone the prosodic and paralinguistic coverage to which we aspire), such manual intervention becomes impractical or too costly.

Thus in the very process of unit-database creation, there is an over-reliance on the phonemic code; moreover, as this reliance is a natural outcome of having to deal with a *textual* input specification of the desired utterance to be synthesized, it is, not surprisingly, usually taken for granted. An alternative point of view is to regard the speech stream in its native, acoustic form, without imposing categorical labels which, it has been argued, originate from grammatical-linguistic concepts bound to the written orthography (Öhman, 2000). This view need not be as controversial as it seems to have been – we certainly do not advocate a senseless abandoning of the rich knowledge accumulated over the long and fruitful history of the phonetic sciences; rather, we seek to overcome some of the inflexibilities of current approaches to speech synthesis, by attempting to more explicitly acknowledge the continuous nature of the acoustic speech stream and thus to relegate the requirement of a phonemic representation to a level higher than that used in current methods.

As outlined in the following sections, this approach leads to an automated process for building a unit-database for a concatenative synthesizer. Specifically, we put aside the phonemic transcription and use only acoustic information both to define (or segment) and to characterize (or label) the individual units. At first glance, this process resembles a form of speech coding, where the acoustically-defined and -encoded segments are free from linguistic constraints. However, as we shall ultimately suggest, a link with the phonemic level may be made, and the leap to a full *text*-to-speech system thereby achieved, via an intermediate representation in terms of quasi-articulatory parameters which may be both estimated from the acoustics and mapped from the input text of the desired utterance.

2. ACOUSTIC SEGMENTATION

To transcend the over-reliance on the phonetic transcription when preparing a unit-database, a key question concerns the type of unit which may be consistently defined using only acoustic information. While numerous types of unit have been proposed in the literature, such as the diphone or "dyad" (Peterson et al., 1958), CV (consonant-vowel) pairs, VCs, CVCs, VCVs, individual phonemes, demi-phones, or other polyphonemic or syllabic combinations, they all rely on careful, *a priori* phonetic segmentation and labeling. By contrast, in the next sections we describe acoustic-prosodic and acoustic-articulatory methods to delimit quasi-syllabic and quasi-articulatory-gestural units automatically.

The speech data used both to illustrate and later to evaluate our methods, are three stories recorded by one female, native speaker of Japanese (Iida et al., 1998). The stories were purposely designed to naturally evoke speaking styles characteristic of anger, joy, and sadness, respectively. In total, there are 1370 utterances, or over 2 hours of speech.

ISBN 1-876346-48-5

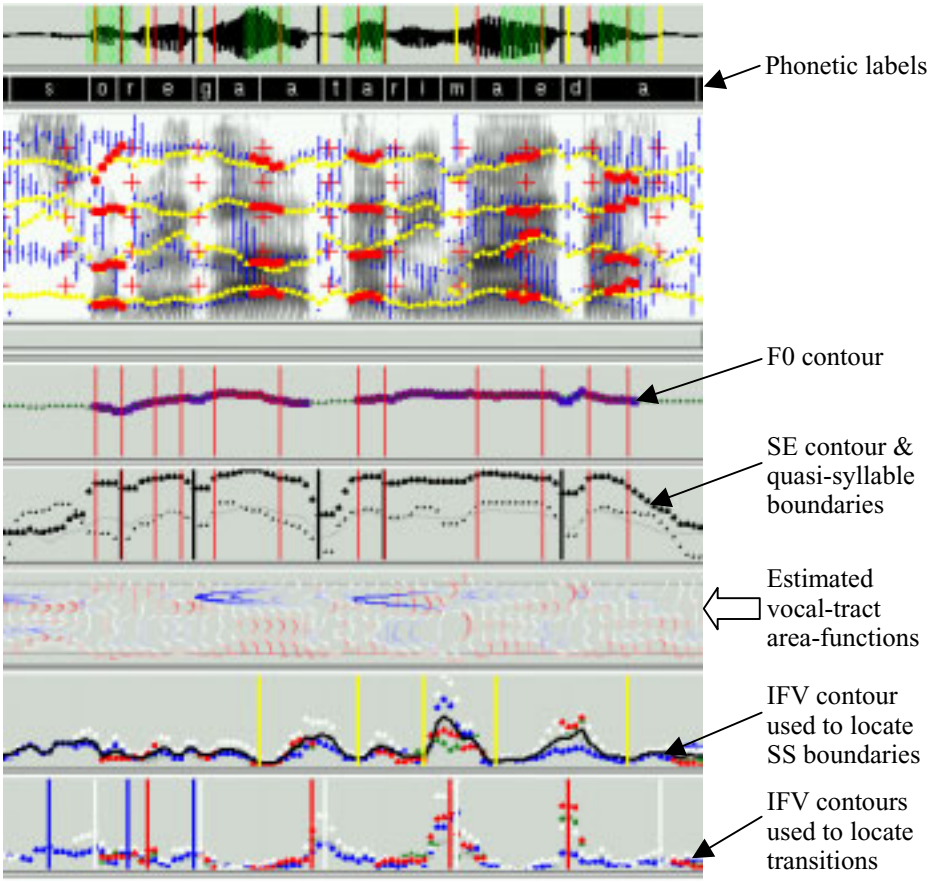

Figure 1. An example illustrating quasi-syllabic and quasi-articulatory-gestural segmentation of continuous speech. The 1.1sec utterance is taken from the angry subset of the data recorded by our female speaker of Japanese.

2.1 Quasi-Syllabic Units

Although 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 definitive specification at either the linguistic or acoustic level. Still, effective methods of quasi-syllabification using acoustic measurements have been known for many years: e.g., Mermelstein (1975) used the convex-hull algorithm to locate significant minima (dips or valleys) in the contour of sonorant-energy (SE) across an utterance. Here, SE is defined as the mean energy within a frequency range which encompasses F0 and roughly the first three formants of voiced sounds while excluding the higher-frequency turbulence-noise of certain obstruents. For our speaker, this frequency range is fixed at 60–3000 Hz; meanwhile, in the convex-hull algorithm we use an amplitude threshold of 2dB and a minimum unit-length of 80msec.

Figure 1 illustrates the automatic quasi-syllabification of a short utterance, taken from the angry subset. The dark data-points in the second panel below the spectrogram show the SE contour which is used to segment the Japanese phrase “sore ga atarimae da” (meaning “that is obvious!”) into the six quasi-syllables “so – re – gaa – ta – rimae – da”, delimited by the dark vertical lines which are duplicated in the speech-waveform panel. The analysis yields CV units, except where there is not a sufficient dip in the SE contour, as in the intervocalic bilabial-nasal in “..rimae..” and in the double-moraic “gaa”. However, these are not regarded as errors – in the context of unit-database creation, we accept the segments as defined by the acoustic characteristics, without imposing any explicit phonological constraints.

2.2 Quasi-Articulatory-Gestural Units

Complementary to the acoustic-prosodic information used above to obtain quasi-syllables, a robust transformation from acoustics to an articulatory representation could yield more physiologically relevant *quasi-articulatory gestures*. While both formant estimation and acoustic-to-articulatory inversion are unresolved research problems, we apply two robust methods to the task of deriving such articulatory gestural units from the acoustics of continuous speech.

2.2.1 Robust acoustic-to-articulatory inversion

Each of the first four formant frequencies and bandwidths are independently estimated at each analysis frame, using a linear transformation of the cepstrum as proposed by Broad and Clermont (1989). The linear regression coefficients which map the formants from the cepstrum are computed using a set of careful measurements in vowel steady-states of our speaker. Comparing the original and the re-estimated values of F1, F2, F3 and F4, we find high correlations (0.89, 0.95, 0.88 and 0.85, respectively). Although formants may properly be defined only when there is an acoustic source to excite vocal-tract resonances, we nevertheless apply the mapping at every frame. As shown by the yellow points on the spectrogram in Fig. 1, the method yields continuous and non-overlapping, quasi-formant trajectories across the entire utterance. Our results confirm Broad and Clermont’s (1989) observation that while the mapping is not precise, nor are there the types of gross errors common to other formant estimation methods.

ISBN 1-876346-48-5

The formant-pattern at every frame is then used to estimate the area-function of the vocal-tract from the glottis to the lips, using a linear-prediction method of inversion which includes closed-glottis correction of formant bandwidths, optimization of the vocal-tract length, and an acoustically-relevant parameterisation of vocal-tract shape (Mokhtari, 1998; Mokhtari and Clermont, 2000). The area-functions thus estimated are shown (for every second frame) in the third panel below the spectrogram in Fig. 1, where the horizontal reference-lines at the bottom and at the top represent respectively the mean position of the glottis (at 0cm) and the lips (at 13.1cm) for our speaker. For greater clarity, the areas are colour-graded such that open cavities (to the left) are blue, and constrictions (to the right) are red. Furthermore, acknowledging longitudinal movements at both the larynx and the lips, all area-functions are aligned at a point mid-way along their entire length (Mokhtari, 1998).

Pending rigorous evaluations, our estimated area-functions can not be claimed to be those produced by the speaker in reality. Indeed, as the inversion and formant estimation methods are applicable ideally to non-nasalised vocoids, it is not surprising to obtain some inexplicable area-functions, such as those at /g/, where a palatal rather than a velar constriction is estimated. Nevertheless, several qualitative observations do tend to uphold the credibility of the area-functions from an articulatory-phonetic point of view: e.g., (i) a palatal constriction and open back-cavity for /i/ in "..rim..", and similar but more neutralized tendencies for /e/ in both "..reg.." and "..aed.."; (ii) a pharyngeal constriction and open front-cavity for the double-moraic /a/ in "..ga-a..", and similar but more neutralized tendencies for /a/ in "..tari.." and in "..mae.."; (iii) a constriction at the lip-end for /m/ in "..ima.."; (iv) an alveolar-like constriction for /d/ in "..eda.."; and (v) almost consistently small areas above the glottis where the usually narrow larynx-tube is expected. Moreover, despite the absence of dynamic constraints, both the length and shape of the area-functions exhibit a smooth evolution across the utterance. In the next section we outline methods to exploit such utterance-length sequences of estimated area-functions, with the aim of locating articulatory states and transitions.

2.2.2 Articulatory states and transitions

To arrive at an articulatorily-motivated segmentation, let us consider the physiological phonetic theory of Peterson and Shoup (1966), who define two types of articulatory states: the *steady-state* (SS) and the *controlled movement* (CM). The speech stream is then produced by a continuum of SSs and CMs, interspersed with faster movements or transitions. In the context of building a unit-database for concatenative synthesis, this articulatory classification suggests at least two distinct types of unit, depending on whether the unit boundaries are located at SSs or at points of transition.

As the physiological phonetic theory does not explicitly define the functions necessary to quantify the positions and velocities of articulators, there are a variety of possible approaches. For example, Broad (1972) showed that the absolute rates of change at different points along estimated area-functions provide complementary information for phonetic segmentation, the points being loosely related to distinct articulators. Inspired by that approach, we define the following 3 regions along the length of centre-aligned area-functions, with the mean position of the glottis at 0 and of the lips at 1: the tongue-body (TB) within [0.3–0.8], the tongue-tip (TT) within [0.7–0.9], and the lips (Lp) within [0.9–1.0]. The activity in each of these articulators can then be quantified by computing an inter-frame variance (IFV) of the (logarithmic) areas within the corresponding region, e.g. in groups of five consecutive area-functions at a time.

The average of the three profiles of activity thus obtained is shown in Fig. 1 by the black curve in the fourth panel below the spectrogram. Articulatory SSs are located as significant minima along that curve using the convex-hull algorithm, and are shown by vertical yellow lines. In this example, they are found to coincide with five vowels, including the centre of the double-moraic /a/ and of the phrase-final /a/, the /a/ part of "..ae..", and the /a/ and /i/ in "..tarim..". A segmentation purely on the basis of these SSs would yield the following sequence of six quasi-articulatory-gestural units: "…sorega – at(a) – ari – ima – aeda – a…".

Alternatively to SSs, *transitions* can be located by using the convex-hull algorithm to find *maxima* in the profile of IFV computed in groups of *delta*-area-functions, which have the desirable property that both SSs and CMs yield relatively low IFV values compared with transitions. By contrast with a SS where *all* articulators are relatively stationary, a transition may involve as few as only one articulator. It is therefore necessary to treat the IFV profiles of the three vocal-tract regions independently. The three colour-coded profiles in the bottom panel of Fig.1 show the amount of (accelerational) activity for TB (blue), TT (white), and Lp (red). While some of the located maxima are articulatorily questionable, many do match their expected type: e.g., the TB in /g/ of "..ega..", the TT in the transition boundary of "..so.." and in /t/ of "..ata..", and the Lp at the bilabial in "..ima..". Imposing a minimum unit-duration of 80 msec, the final boundaries (yellow vertical lines in the waveform panel) yield the following seven units: "…sor(e) – e(g) – gaa – tarim – mae(d) – da…". It is interesting to note that the quasi-articulatory boundaries in /g/, /t/ and /d/ are quite close to the quasi-syllabic boundaries found earlier using only acoustic-prosodic information.

3. UNIT SELECTION FOR SYNTHESIS

On the basis of our description of quasi-syllabic (QS) and quasi-articulatory-gestural (QAG) units in the preceding section, the following three, basic types of unit are defined: quasi-syllabic units (**QS**); QAG units delimited by SSs (**QAG1**); and QAG units delimited by transitions (**QAG2**). In addition, boundaries of different unit-types are combined while adhering to a minimum unit-length of 80msec, to form the following three, compound types: QS units further divided at QAG1 boundaries (**QSAG1**); QS units further divided at QAG2 boundaries (**QSAG2**); QAG1-SSs combined with QAG2-transitions (**QAG12**).

In lieu of the phonetic labels, each unit is characterized by

 ISBN 1-876346-48-5

Phonetic labels of original unit	Phonetic labels of selected units	Percent exact match
sil (9841)	sil (8902), sil-sil (451), a-sil (89), sil-s (48), U-sil (26), tt (26), sil-k (17)	90.5
a-sil (1853)	a-sil (1408), sil (168), a-tt (58), a (42), o-sil (21), a-k (14), a-t (12)	76.0
o-sil (814)	o-sil (574), a-sil (44), o-k (29), o-o-sil (25), o-t (14), o-o-t (10), o-o-k (10)	70.5
i-sil (697)	i-sil (491), sil (55), i-t (29), e-sil (16), i-tt (14), i (11), i-k (6)	70.4
e-sil (627)	e-sil (445), sil (64), a-tt (15), u-sil (10), i-sil (9), e-t (8), e-e-sil (7)	71.0
a-N (467)	a-N (297), a-n (59), a-m (23), a-N-n (8), a-i (7), o-N (5), a-n-o (5)	63.6
a-tt (459)	a-tt (275), a-sil (39), a-t (34), a-k (19), u-sil (11), sil (10), e-sil (10)	59.9
a-sh (407)	a-sh (267), a-sh-I (42), a-s (34), o-sh (10), a-ssh (5), e-sh (4), a-ch-I (4)	65.6
sil-sil (390)	sil (345), sil-sil (29), a-sil (3), sil-sh (2), i-sil (2), tt (1), sil-s (1)	7.4
u-sil (356)	u-sil (212), sil (26), e-sil (15), o-sil (10), o-k (8), a-tt (8), a-sil (7)	59.6
sil-s (345)	sil-s (147), sil (113), sil-sh (28), sil-k (12), sil-ts (8), sil-sil (6), sil-f (5)	42.6
e-N (332)	e-N (203), e-n (15), i-N (14), e-e-n (14), e-m (10), e-N-n (7), e-o (6)	61.1

Table 1. The 12 most frequently-occurring units (and their total number shown in parentheses) yielded by the QAG1 method of segmentation (quasi-articulatory gestures delimited by steady-states), according to an *a posteriori* analysis using the phonetic labels. Also listed for each original unit, are the 7 most frequently-occurring units selected on the basis of an acoustic distance (as described in section 3); and the percentage of exact phonetic-label matches.

its duration, and by the mean and first 4 discrete-cosine-transform (DCT) coefficients of its contours of F1, F2, F3, F4, F0 (interpolated through unvoiced regions), SE, and a higher-frequency energy (HFE) measured in the frequency band 3400–6000 Hz. A concatenative speech-to-speech synthesizer is then tested by holding out each of the 1370 utterances in turn, and searching for the closest match to each of its units from among the units in the remaining 1369 utterances. For this purpose, we use a Euclidean distance on all of the acoustic parameters listed above, where each parameter is weighted by the reciprocal of its standard-deviation across the entire dataset.

While phonetic labels were completely disregarded in the automatic segmentation process, they now provide one way to evaluate the segmentation and unit-selection methods. In particular, the labels spanned by each unit of an original utterance are compared with the labels of the corresponding selected unit. One objective measure of performance is then the proportion of exactly-matching labels. Ignoring all units labeled as silence and all labels which occur only once in the entire data, the proportion of exact matches for each segmentation method is as follows: 25.3% (QS), 38.2% (QAG1), 21.4% (QAG2), 31.7% (QSAG1), 24.0% (QSAG2), and 32.7% (QAG12). Although these figures give only a conservative indication of resynthesis accuracy, it is interesting to note that the highest score is obtained for the QAG1 method where units are delimited at articulatory steady-states and thereby capture the intervening dynamics of CMs and transitions.

An indication of the types of errors in the QAG1 method is given in Table 1. A glance at the most common mismatches reveals that many of them involve label-related additions, deletions or substitutions which are acoustically, and perhaps perceptually, of little consequence. Indeed, in reading the table it is important to bear in mind that each unit extends only to the SS of the labels at either end; e.g., "a-tt" which extends only to the silence gap prior to the stop-burst, is acoustically interchangeable with "a-sil". The unit-selection accuracy is therefore much higher than the quoted figures suggest. On the other hand, certain phonetic confusions such as among classes of nasals or fricatives, do call for improvements in both segmentation and acoustic parameterisation methods. Perceptual experiments are under way, to formally assess the auditory intelligibility and naturalness of the resynthesised speech. Ultimately, we plan to adopt an intermediate articulatory representation which can be mapped both from text and from acoustics, and which will therefore lead to a full text-to-speech system based on the concepts expounded in this paper.

ACKNOWLEDGMENTS

This work is supported by the Japan Science and Technology (JST) Corporation under CREST Project 131. We would like to express our gratitude to all members of the ESP Project at ATR, especially Carlos Toshinori Ishi for his suggestions and critical discussions.

REFERENCES

[1] D. J. Broad, "Formants in automatic speech recognition," *Int. J. Man-Machine Studies*, vol. 4, pp. 411-424, 1972.

[2] D. J. Broad and F. Clermont, "Formant estimation by linear transformation of the LPC cepstrum," *J. Acoust. Soc. Am.*, vol. 86, pp. 2013-2017, 1989.

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

[4] P. Mermelstein, "Automatic segmentation of speech into syllabic units," *J. Acoust. Soc. Am.*, vol. 58, pp. 880-883, 1975.

[5] P. Mokhtari, "An acoustic-phonetic and articulatory study of speech-speaker dichotomy," Doctoral Thesis, The University of New South Wales, Australia, 1998.

[6] P. Mokhtari and F. Clermont, "New perspectives on linear-prediction modelling of the vocal-tract: uniqueness, formant-dependence and shape parameterisation," *Proc. 8th Australian Int. Conf. on Speech Science and Tech.*, Canberra, pp. 478-483, 2000.

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

[8] G. E. Peterson and J. E. Shoup, "A physiological theory of phonetics," *J. Speech Hear. Res.*, vol. 9, pp. 5-67, 1966.

[9] G E. Peterson, W. S.-Y. Wang and E. Sivertsen, "Segmentation techniques in speech synthesis," *J. Acoust. Soc. Am.*, vol. 30, pp. 739-742, 1958.