1—6—16 A proposal for acoustic-articulatory gestural units in concatenative speech synthesis

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

Faced with the problem of dealing with a text-based input specification (whether orthographic or phonemic), most methods of concatenative text-to-speech synthesis use basic speech units which are defined linguistically. Naturally, such linguistically-defined units as diphones, demiphones, CVs, VCs, syllables, or even phonemes themselves, facilitate a direct mapping from the input textual specification to the individual sound units required to synthesise the spoken utterance. However, such a constrained specification also presupposes a sufficiently accurate, phonetic segmentation and labelling of the spoken data; consequently, creation or extension of a unit-database, especially one which is sufficiently large and rich to allow synthesis of natural, expressive, conversational speech, is non-trivial, labour-intensive, and costly. Nevertheless, when such costs are borne, concatenative speech synthesis does afford a high degree of perceived naturalness, owing mainly to the preservation of the many naturally-occurring acoustic variabilities such as those due to coarticulation and speaker characteristics.

In a first attempt to transcend the over-reliance on the written or linguistic code while still taking advantage of the benefits of concatenative synthesis, we have recently proposed a novel method wherein *quasi-syllabic* units of speech are defined, characterised, and selected for concatenation, purely on the basis of *acoustic* information (Mokhtari & Campbell, 2002). As therein noted, and more recently elaborated (Campbell & Mokhtari, 2003), the problem of dealing with a textual input specification of the desired utterance might then be overcome by using either a conventional concatenative method, or indeed a more traditional, rule-based method of text-to-speech, in order to first generate at least a phonetically acceptable, acoustic rendition of the utterance. That intermediate acoustic output could then be subjected to the same, automatic methods of quasi-syllabification and unit characterisation proposed earlier, and each syllabic unit thereby replaced with acoustically-matching units chosen from a large database of natural speech. An alternative, and perhaps scientifically more rewarding, approach would be to use an intermediate representation of the utterance in the articulatory domain – specifically, an articulatory representation which might be mapped (however approximately) both from the text domain and from the acoustic domain.

While the advantages of an intermediate articulatory representation of speech have long been heralded in the context of both synthesis and recognition, such representations, particularly those estimated from the acoustics by methods of speech inversion, remain underexploited. Previous work has shown, however, that acoustically-estimated vocal-tract area-functions can be used to explain and to model phonetic variability (Mokhtari & Tanaka, 2000), inter-speaker variability (Mokhtari, Clermont & Tanaka, 2000), and emotion-related variability in speech (Mokhtari, Iida & Campbell, 2001); and there has been a recent proposal to more explicitly use articulatory modelling in concatenative speech synthesis (Sondhi, 2002). In this paper we extend our earlier proposal of acoustic selection of quasi-syllabic units, by outlining a new method of synthesis whereby vocal-tract area-functions are estimated from the acoustics of speech and are then explicitly used to delimit, and potentially to characterise, the basic units for concatenation.

2 Why resort to the articulatory domain?

Our motivations for expending the extra conceptual and computational effort of mapping from acoustics to the articulatory domain are well-founded.

Firstly, as mentioned above, an articulatory-domain specification may facilitate the transition from our current *speech*-to-speech system to a more complete *text*-to-speech conversion, if only because "the relation between articulation and phoneme is more nearly one-to-one than that between phoneme and sound" (Liberman et al., 1959, p.1496). In particular, an input phonetic string may be converted to a sequence of vocal-tract area-functions, for example via components of Browman & Goldstein's (1990) gestural computational model; and then these area-functions could be analysed using the same methods that we shall describe below for treating acoustically-estimated area-functions, with the aim of obtaining a physiologically meaningful segmentation of the utterance into basic units which could then be retrieved from an acoustic unit-database.

Secondly, as just alluded to, the physiologically-meaningful information conveyed in an articulatory-domain representation of speech can be expected to provide a more effective (or a more optimal) segmentation of the continuous, acoustic speech stream. Indeed, the original definition of the "dyad" – the original term for what is today regarded as a "diphone" and used extensively as a basic unit for synthesis – referred to an "articulatory sequence pair" involving "all supra-laryngeal phenomena in the vocal tract" (Peterson et al., 1958, p.740). As noted in that seminal study, "if synthesized speech is to sound natural, the normal dynamics of speech production must be maintained"; hence, the boundaries of a dyad, and therefore the junctions of adjacent units, are defined "at relatively sustained positions" in the speech stream. Referring back to the linguistic code, these positions may coincide with either voiceless or voiced fricatives, silent gaps of plosives, or vocalic steady-states. However, disregarding the orthographic transcription of an utterance and instead considering only the information that is present in the acoustic speech signal, the phonetic identity of units whose boundaries are defined on the basis of "relatively sustained positions" will depend on all the factors which contribute to the particular speaking style of the utterance, and the units therefore need not conform with a strict linguistic definition of a "dyad" as gleaned from an analysis of the utterance text.

In lieu of the phonetic-linguistic "diphone", an

調音動作単位による波形接続型音声合成の提案
○ パーハム モクタリ、 ニック キャンベル
JST/CREST ESP Project, ATR-HIS Labs, Kyoto.

acoustic-articulatory analysis motivated above and described in the following section may yield *quasi-articulatory gestures* which span the interval from one relatively sustained articulatory state, through one or more relatively transitory gestures or controlled movements (Peterson & Shoup, 1966), to the next significantly sustained position.

3 Quasi-articulatory gestures from acoustics

As our acoustic-articulatory analyses are intended to be applied in an unsupervised fashion to very large amounts of natural, recorded speech, particular emphasis is placed on robustness, even at the expense of precision. Accordingly, the first step in obtaining an articulatory representation is to estimate the first four formant frequencies and bandwidths, by linear transformation of the linear-prediction (LP) cepstrum, as proposed by Broad & Clermont (1989). While formants are properly defined only in voiced segments of speech and the cepstrum-to-formant mapping is trained on a balanced set of vowel steady-states of our selected speaker, continuous quasi-formant contours are obtained by linear transformations of the cepstrum at every analysis frame. The formants are then used, independently at each frame, to estimate the length and shape of the vocal-tract by an LP-based method of inversion together with a parameterisation of the vocal-tract shape in terms of the first four, odd-indexed cosine and sine coefficients (Mokhtari & Clermont, 2000). An arbitrary but relatively unbiased anchor-point is then defined at the mid-length of each area-function, and the frame-wise area-functions are thus centre-aligned (Mokhtari, 1998) under the assumption of a continuously varying glottal-height and lip-protrusion.

Previous analyses of the vowel steady-states of our female speaker had revealed an average estimated vocal-tract length of 13.1cm (Mokhtari, Iida & Campbell, 2001). This average length, which may be associated approximately with the vocal-tract anatomical structure, is here subdivided into five regions (specified in terms of the ratio along the length of the vocal-tract from the glottis at 0 to the lips at 1); these regions correspond grossly to the following, primary articulators or structures: larynx tube [0.0 – 0.1], lower pharynx [0.1 – 0.3], tongue body [0.3 – 0.8], tongue tip [0.7 – 0.9], and lips [0.9 – 1.0]. For each of these five regions in turn, a contour of articulatory variability is obtained across the entire utterance by computing the variance in groups of five consecutive area-functions, advancing one frame at a time.

A composite contour can then be found by averaging the five individual contours, and the convex-hull algorithm then used to locate the significant dips or valleys which correspond to the locally invariant or relatively sustained positions. However, a more interesting and potentially more revealing analysis is to locate the significant minima in each of the individual contours in turn, and then to combine all the resulting segmentation boundaries. Such an approach was indeed implicated by Broad (1972) in the context of using the formants to indirectly estimate articulatory states that may delimit phonemes, by separately detecting locations of variability around for example the tongue dorsum region versus the lip region. In the present work, however, we propose to explicitly use vocal-tract area-functions estimated by a more complete method of inversion, and to computationally integrate the quasi-articulatory variability within vocal-tract regions which roughly correpond to major physical structures or articulators. Rather than an acoustic-phonetic segmentation as aspired in Broad's (1972) classic work, our segmentation therefore yields units which may be regarded as *quasi-articulatory gestures*.

4 Preliminary results and ongoing work

Automatic segmentation of one database of recorded, expressive speech (Iida et al., 1998) yielded more than 60,000 quasi-articulatory gestural units, with a mean duration of about 150msec. Compared with our previous automatic segmentation of the same database into *quasi-syllabic* units (Mokhtari & Campbell, 2002), the present analysis yielded about a 50% increase in the number of units, whose durations are therefore typically shorter. While this result was expected, it remains to be seen whether the shorter durations and the greater occurrence of these new, quasi-articulatory units will help to overcome the problem of acoustic-phonetic coverage of the resulting unit-database. In particular, a preliminary informal assessment of the efficacy of using such units in speech-to-speech synthesis (where the utterance to be synthesised is held out from the available unit-data) indicates that while the new method is indeed able to find a greater range of acoustic-phonetically similar units from which to select, there is also an increased amount of audible distortion at join boundaries. In ongoing research, we are aiming to find a compromise between our previous, prosodically-defined quasi-syllabic units (whose boundaries are by definition low in sonorant energy, and therefore easier to join without distortions), and the present, articulatorily-motivated segmentation. Synthesised speech samples will be demonstrated, together with a more complete analysis of the acoustic-phonetic and estimated articulatory properties of the gestural units.

Acknowledgements

We are grateful to all members of the JST-CREST ESP project, especially Carlos Toshinori Ishi, for moral and scientific support and encouragements.

References

Broad, D.J. (1972). "Formants in automatic speech recognition", *Int. J. Man-Macghine Studies* 4, 411-424.

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

Browman, C.P. & Goldstein, L. (1990). "Gestural specification using dynamically-defined articulatory structures", *J. Phonetics* 18, 299-320.

Campbell, N. & Mokhtari, P. (2003). "Using a non-spontaneous speech synthesiser as a driver for a spontaneous speech synthesiser", in *Proc. ISCA & IEEE Workshop on Spontan. Speech Process. and Recog.* (SSPR), Tokyo.

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.

Liberman, A.M., Ingemann, F., Lisker, L., Delattre, P. & Cooper, F.S. (1959). "Minimal rules for synthesizing speech", *J. Acoust. Soc. Am.* 31, 1490-1499.

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

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

Mokhtari, P. & Campbell, N. (2002). "Automatic characterisation of quasi-syllabic units for speech synthesis based on acoustic parameter trajectories: a proposal and first results", in *Proc. Autumn-02 Meet. of the Acoust. Soc. Japan*, Akita, 233-234.

Mokhtari, P. & Clermont, F. (2000). "New perspectives on linear-prediction modelling of the vocal-tract: uniqueness, formant-dependence and shape parameterisation", in *Proc. 8th Australian Internat. Conf. on Speech Science and Technology* (SST), Canberra, 478-483.

Mokhtari, P., Clermont, F. & Tanaka, K. (2000). "Toward an acoustic-articuloatory model of inter-speaker variability", in *Proc. 6th Internat. Conf. on Spoken Lang. Process.* (ICSLP), Beijing, Vol. II, 158-161.

Mokhtari, P., Iida, A. & Campbell, N. (2001). "Some articulatory correlates of emotion variability in speech: a preliminary study on spoken Japanese vowels", in *Proc. Internat. Conf. on Speech Process.* (ICSP), Taejon, 431-436.

Mokhtari, P. & Tanaka, K. (2000). "Principal components of estimated vocal-tract shapes of Japanese vowels", in *Proc. Spring-00 Meet. Of the Acoust. Soc. Japan*, Funabashi, 327-328.

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.

Peterson, G.E. & Shoup, J. E. (1966). "A physiological theory of phonetics", *J. Speech Hear. Res.* 9, 5-.

Peterson, G.E., Wang, W.S-Y. & Sivertsen, E. (1958). "Segmentation techniques in speech synthesis", *J. Acoust. Soc. Am.* 30, 739-742.

Sondhi, M. M. (2002). Articulatory modeling : a possible role in concatenative text-to-speech synthesis, in *Proc. IEEE Workshop on Speech Synthesis*, Santa Monica.