Automatic Detection of Acoustic Centres of Reliability for Tagging Paralinguistic Information in Expressive Speech

Parham Mokhtari and Nick Campbell
JST-CREST / ATR Human Information Science Laboratories, Kyoto, Japan
parham@atr.co.jp nick@atr.co.jp

Abstract

Preparation of a unit-database to be used in concatenative speech synthesis demands sufficiently robust, unsupervised algorithms for processing the typically huge corpora. The demands are even more stringent when considering a corpus large enough to capture a wide variety of speaking-styles and emotions, even of a single speaker. This paper describes a method of combining robust acoustic-prosodic and cepstral analyses to locate centres of acoustic-phonetic reliability in the speech stream, wherein physiologically meaningful parameters related to voice quality can be estimated more reliably. These parameters which describe the state of glottal phonation and of supralaryngeal articulation, can then provide a paralinguistic annotation of the unit-database, thereby enabling speech synthesis with a greater variety of expressions and speaking-styles.

1. Introduction

Concatenative speech synthesis technology based on unit-selection relies crucially on the creation, maintenance, and mark-up of spoken language data. In the framework of the Japan Science and Technology (JST) Corporation's project on Expressive Speech Processing (ESP), our approach is to use very large corpora of speech recorded in a variety of speaking-styles and natural, expressive situations. In preparing a unit-database to be used for synthesis, it is therefore imperative to develop sufficiently unsupervised methods for reliable and robust processing, labelling, and annotation of the recorded speech data.

In this paper we leave aside issues related to the phonetic segmentation and labelling, and focus on the relatively more neglected issue of paralinguistic annotation. Indeed, to gain the flexibility of synthesising speech in a variety of speaking-styles and emotions, the challenge is to augment the basic, phonetic and phonologic-prosodic labels with an additional layer (or layers) of paralinguistic tags, which will allow a more judicious selection of units appropriate to a desired speaking-style or expressive quality. While the ultimate test of the efficacy of paralinguistic tags is perhaps a perceptual evaluation of the synthesised speech, we believe that the annotation process itself must pay heed to the fact that the origin of both phonetic and expressive variability in speech lies in the production domain. Our approach is therefore to map the measurable acoustics of speech back to physiologically-related parameters; to compute principal axes of emotion-related variation in those speech-production parameters; and then to test the perceptual relevance of (and perhaps to refine) those axes of variation by auditory evaluations.

While the problem of mapping physiological parameters from acoustics – the well-known speech inverse problem – has been studied for forty years or more and there is still not a general consensus on any single, best solution, the formants (or resonances of the vocal-tract) still hold a place of pride in that they are arguably the acoustic parameters most closely interpretable in articulatory terms. The formants do indeed play a central role in many of the classic and more recent methods of speech inversion (e.g., Schroeder, 1967; McGowan, 1994; Yehia & Itakura, 1996; Schoentgen & Ciocea, 1997). Furthermore, and despite the inevitable circularity of the argument, it is well-known that the most reliable estimates of glottal parameters from either the speech waveform or the speech spectrum may be obtained by methods which rely (whether explicitly or indirectly) on good estimates of the formants.

However, it is also well-known that it is difficult to measure formants both automatically and reliably, whether only in voiced segments of continuous speech or in more controlled vocalic steady-states; hence the need for supervision and manual correction whenever formants are required in any serious study. In view of the very large amounts of recorded speech data required in our efforts to build a unit-database suitable for truly expressive speech synthesis, the challenge is therefore to develop completely unsupervised algorithms for accessing the formants wherever they may be most reliably measured in natural, spontaneous speech. With these motivations, in the following section we consider in greater detail the concept of reliable centres in speech, and how they may be located automatically in very large corpora.

2. Centres of Reliability

In this section we first outline a rationale, then describe our computational methods for locating centres of reliability in the acoustic speech stream.

2.1. A Rationale

In the research context introduced above and in agreement with Ohman's (1967) thesis, we view the speech stream in its most primitive form as a continuum of vocoids and contoids. While contoids can be regarded as modulations or articulatorily constrictive perturbations in the speech stream, vocoids are the more likely candidates for syllabic nuclei. Moreover, as vocoids generally exhibit relatively stable spectral characteristics and relatively high levels of acoustic energy, they are also the more likely candidates for what Lea et al. (1975; 1980) have termed "islands of reliability". It is indeed within just such portions of the acoustic speech stream that the perceptually and articulatorily salient, formant parameters can be measured with the greatest reliability; and by extension, these centres of reliability ought to yield the most reliable estimates of the types of voice-quality settings required to produce a paralinguistic annotation of recorded speech.

Inspired by these considerations, we advance a definition of "reliable centre" in terms of the following, three main criteria: it must (i) lie inside a syllabic nucleus; (ii) have stable spectral characteristics; and (iii) be able to yield good initial estimates of the formants. Expanding on each of these criteria in acoustic terms, our rationale for locating a centre of reliability prescribes: (i) a vocoid or fully-voiced region with relatively high sonorant energy;

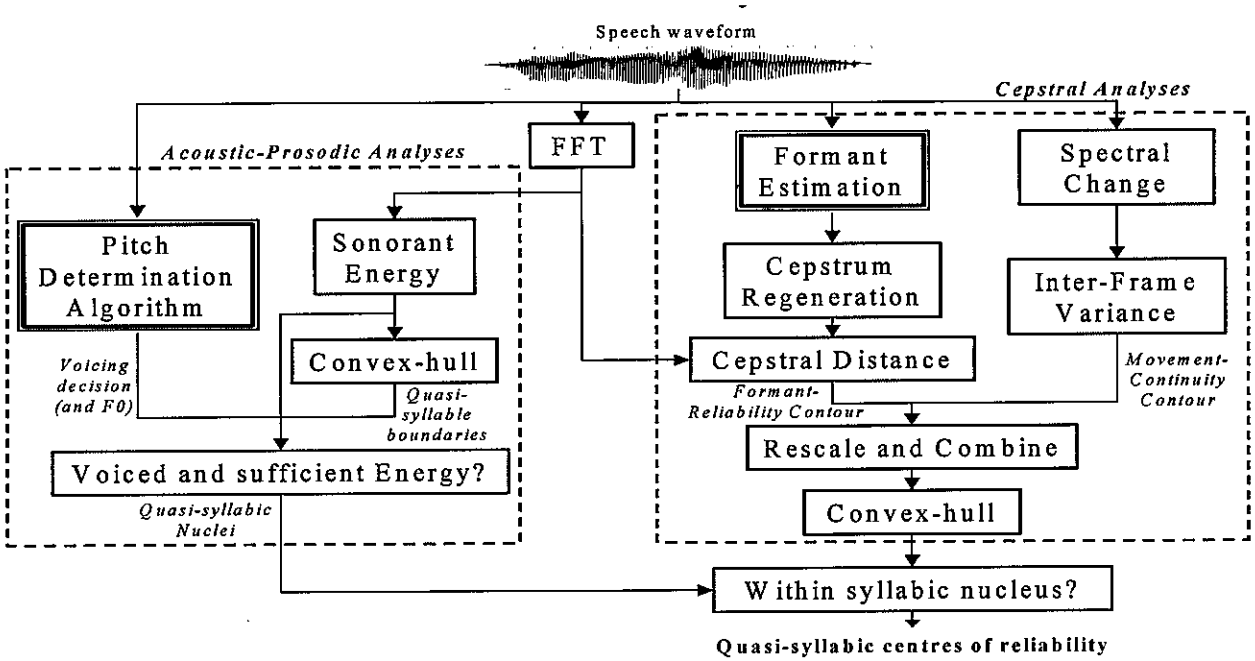

Figure 1. Flow-chart of our completely unsupervised algorithm to locate centres of reliability in the acoustic speech stream (see section 2.2 for detailed discussion).

(ii) either a spectral steady-state or a region with relatively smooth spectral-change; and (iii) a region where sufficiently robust, initial formant estimates provide a relatively close match with respect to the original acoustic data, e.g. in a spectral-matching sense. In the next section we describe in detail our algorithm which follows directly from the above rationale.

2.2. An Unsupervised Algorithm

As shown in Figure 1, our algorithm proceeds along two conceptually parallel strands: one dealing with *acoustic-prosodics*, and the other reliant mainly on *cepstral* analyses. The input speech utterance may in principle be of any length (e.g., word-, phrase-, sentence-length or longer) and may also contain pauses of any duration. The results of the two strands of analysis are finally combined to yield an estimate of the reliable centres, as described below and illustrated with an example in Figure 2.

The acoustic-prosodic strand is dominated by the quasi-syllabic segmentation method proposed by Mermelstein (1975), wherein the so-called convex-hull algorithm is used to detect significant dips (or valleys) in the time-contour of *sonorant energy* (see the second panel below the spectrogram in Figure 2). The latter is defined as the acoustic energy (in dB) within the frequency band from around 60 Hz to 3000 Hz which, after some gender- (or speaker-) specific adjustments, is intended to encompass the sonorant range from the fundamental frequency F0 up to approximately the third formant F3, while largely excluding the higher-frequency energies associated mainly with turbulent noise produced in many classes of contoids. Within each quasi-syllabic segment thus found, the boundaries of the corresponding, quasi-syllabic *nucleus* are then located by starting at the peak in sonorant-energy and extending frame by frame both to the left and to the right, so long as the frame is voiced (as estimated by a waveform-correlation threshold in conjuction with a pitch-determination algorithm based on the sub-harmonic summation method of Hermes, 1988) and the sonorant energy also remains above a certain threshold (around 0.8 of its range within the quasi-syllable); this centre-outward processing was inspired by Lea & Clermont (1984).

Meanwhile in the strand dominated by cepstral analyses, Peterson and Shoup's (1966) articulatory-phonetic concepts of a steady-state and of a continuous movement are recast in acoustic terms (albeit crudely, given that the mapping between articulator movements and the resulting acoustics is generally non-linear), to obtain a time-contour of *spectral movement-continuity*. This is achieved by first obtaining a measure of spectral change, as afforded simply by the delta-cepstrum; the local (dis)continuity in spectral change is then quantified using a cepstral distance measure to compute an inter-frame variance (Mokhtari, 1998) in every group of five consecutive frames of delta-cepstra. The lower the value of this variance, the smoother or more continuous the local change in spectral characteristics, whether a steady-state with almost no change, or a smoothly changing dynamic segment.

The linear-prediction (LP) cepstra computed for the above analyses are also used to obtain an initial estimate of the first four formant frequencies and bandwidths independently for every analysis frame, using the linear cepstrum-to-formant mapping first proposed by Broad & Clermont (1989) and later pursued by Bayya & Hermansky (1990) and by Högberg (1997). Particularly in the single-speaker (or speaker-dependent) case, Broad & Clermont (1989) have noted the remarkable robustness of the linear mapping from LP-cepstrum to formants, in the sense that although the estimated formants are not highly accurate, nor have they been observed to be grossly incorrect (as can often occur in conventional formant estimation methods based on spectral-matching criteria, where formants can be missed or assigned to the wrong

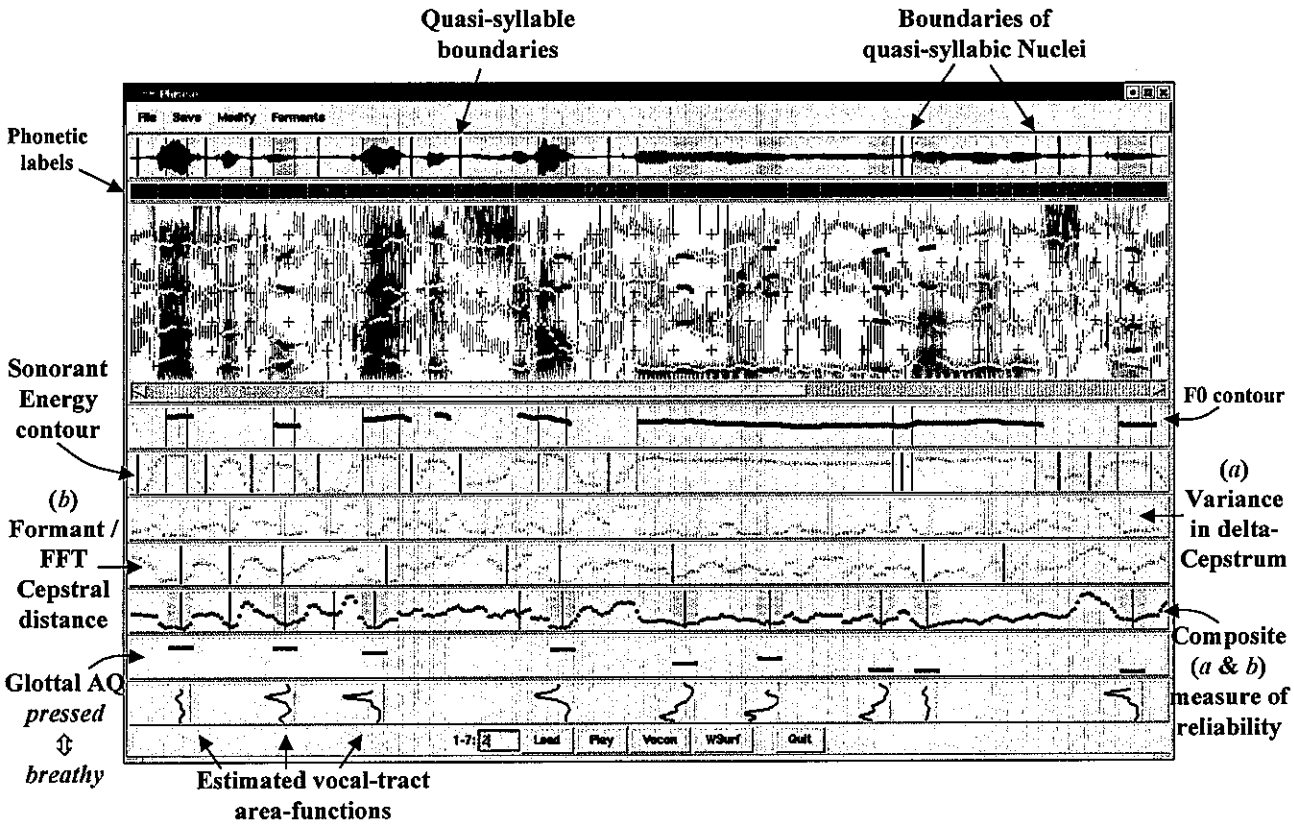

Figure 2. Snapshot of our graphical user interface, implemented in Tcl/Tk and using certain useful functions such as waveform and spectrogram display provided in the Snack package. See text for detailed discussion.

spectral peak, even when dynamic constraints are used). It is precisely this type of robustness which is desirable for unsupervised analyses of very large speech corpora, and which we here find indispensable. Those estimated formants for each frame are then used to regenerate simplified LP-cepstra, which are in turn compared with the corresponding FFT-cepstra computed from the original speech waveform. That comparison using a cepstral distance measure yields a contour of initial formant-(un)reliability, returning for this purpose to the conventional spectral-matching paradigm: the lower the distance value, the more closely matched are the estimated formants with respect to the raw spectral representation.

The two, independent time-contours obtained by the cepstral analyses described above – the spectral-movement continuity contour (see panel "*a*" in Figure 2) and the initial-formant reliability contour (panel "*b*" in Figure 2) – are each linearly rescaled to the range [0,1] and combined simply by averaging pairs of corresponding frames, to yield a composite contour which can be regarded as a measure of the local reliability *and* spectral-change invariance. This composite contour is then subjected to the convex-hull algorithm used earlier in the acoustic-prosodic strand, in order to locate the significant valleys or dips, which signify regions of both formant reliability and spectral-change invariance (shown by the vertical lines superimposed on the composite "*a* & *b*" panel in Figure 2). Finally, only those significant dips of the composite contour are retained which also lie within the boundaries of the quasi-syllabic nuclei found earlier in the prosodic analysis. These retained locations are referred to as *quasi-syllabic centres of reliability*, and the formants estimated at the five consecutive frames around each centre are retained for further, voice quality analysis.

3. Preliminary Results

The algorithm motivated and described in the previous sections was applied to a database of emotional speech recorded by an adult, female, native speaker of Japanese (Iida et al., 1998). Each of the three, read stories were designed to naturally evoke the emotions Anger, Joy, and Sadness, respectively; and each contained more than 400 sentence-length utterances (or more than 30,000 phonemes) stored in separate speech-wave files for independent processing.

Table 1 shows the phonetic distribution of the automatically-located centres of reliability, listing the number of times that a reliable centre happened to coincide with a segment labelled as either one of the five Japanese vowels, or as any other phoneme. As our algorithm makes no prior use of phonetic labels and is therefore independent of phonetic segmentation and labelling, it is encouraging to note that of the total 22175 centres detected across the entire database, only 1501 (or 6.8%) fall into the "other" category, the distribution of which is as follows: n (475), m (276), N (229), w (225), y (112), r (83), d (32), g (23), label undetermined (15), silence (8), h (6), b (5), z (5), j (4), k (1), t (1), sh (1). Clearly, of the non-vowel phonemes located, the nasals are the most common, followed by the liquids and semi-vowels. Of the five vowels themselves, the majority coinciding with centres of reliability were found to be /a/ and /o/ with around 6000 of each, followed by /e/

	ANGRY	JOYFUL	SAD	Total
a	2138	2336	2089	6563
i	861	1219	889	2969
u	614	626	585	1825
e	1244	1242	938	3424
o	2015	2059	1819	5893
other	474	532	495	1501
Total	7346	8014	6815	22175

Table 1. Phonetic distribution of the centres of reliability automatically located in our emotional speech database.

and /i/ with around 3000 of each, and finally /u/ represented by just under 2000 reliable centres.

As can be seen in the example snapshot shown in Figure 2, there is both much to commend and room to improve the performance of the automatic procedure. Fully-voiced and high-sonorant-energy quasi-syllabic nuclei are found in an acoustically consistent way; and in the 2.7sec interval of speech shown, a total of about 8 quasi-syllables and 9 centres of reliability are found. Evidence of the fact that in this preliminary experiment we have chosen to set the various parameters of the algorithm to rather conservative values, can be found for example in the second and the fifth quasi-syllables where, despite the relatively high plateau of sonorant energy, the pitch-detection algorithm (or more correctly its post-processing clean-up of the F0 contour) failed to report the presence of voicing in those syllabic nuclei, hence also leading to a failure to select the candidate centre of reliability in the second syllable. Also, the detected seventh and eighth syllabic nuclei appear to have durations longer than an expert linguist might decide, the former spanning the portion of the utterance labelled as /eNmeeniyu/ and the latter spanning /aneru/. However, while in English these might be divided into 3 or 4, and 2 or 3 syllables respectively, and even in the native Japanese the number of *mora* in each sample might normally be comparable to or even greater than those numbers, acoustic consistency is maintained in that the sonorant-energy contour indeed exhibits very small (if any) dips within those intervals. It is also interesting to note that in the very short, ninth detected syllable, the apparently (and commonly) devoiced vowel in "shik" is nevertheless syllabified by virtue of the plateau in sonorant energy around the high second- and third-formant region.

4. Ongoing Research

As discussed in the introduction, the principal motivations of our present work stem from the application to concatenative speech synthesis, where a fully-automated paralinguistic annotation of a unit-database would allow better selection of units according to the desired speaking-style or expressive quality. The formants which are measured by our algorithm in the most reliable portions of the speech stream, bear one of the most potent keys to unlocking the voice-quality settings related to vocal-tract and glottal characteristics. In the lower panels of Figure 2 are shown the results of two such analyses: the vocal-tract area-function and the glottal amplitude-quotient (AQ), estimated at reliable centres by methods of inversion (Mokhtari, 1998) and inverse-filtering, respectively.

As proposed by Alku & Vilkman (1996) and more extensively validated in our own recent work (Mokhtari & Campbell, 2002a), the AQ parameter which is based on amplitude-domain properties of the estimated glottal-flow waveform, quantifies rather well the auditory impression of voice-quality along the pressed-breathy continuum (Sundberg, 1987), which at least in English has been described as paralinguistically signaling anger at one extreme, and, amongst other things, intimacy at the other (e.g., Laver, 1980). Recent experiments using Japanese data confirm the emotion-related trends in AQ (Mokhtari & Campbell, 2002b), which we therefore plan to use in selecting speech units to be concatenated during synthesis. Finally, we are also in the process of similarly using features of estimated area-functions (shown in the bottom panel of Figure 2), after appropriate reparameterisation in the speaker's emotion-related space of articulatory settings (cf. Mokhtari et al., 2001), to detect features such as consistent lip-pouting or or other tongue-postures during production of natural, expressive speech.

References

Alku, P. & Vilkman, E. (1996). "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering", *Speech Comm.* **18** (2), 131-138.

Bayya, A. & Hermansky, H. (1990). "Towards feature-based speech metric", in *Proc. IEEE Int. Conf. on Acoust., Speech, and Sig. Process.*, 781-784.

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

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

Högberg, J. (1997). "Prediction of formant frequencies from linear combinations of filterbank and cepstral coefficients", KTH-STL-QPSR, Royal Inst. of Tech. Stockholm, Sweden, 41-49.

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.

Laver, J. (1980). *The phonetic description of voice quality*, CPU, Cambridge.

Lea, W. A. (1980). "Prosodic aids to speech recognition", in Lea, W.A. (ed.), *Trends in speech recognition*, Prentice-Hall, New Jersey, 166-205.

Lea, W. A. & Clermont, F. (1984). "Algorithms for acoustic prosodic analysis", in *Proc. IEEE Int. Conf. on Acoust., Speech, and Sig. Process.*, 42.7.1-42.7.4.

Lea, W. A., Medress, M. F. & Skinner, T. E. (1975). "A prosodically guided speech understanding strategy", *IEEE Trans. on Acoust, Speech, and Sig. Process* **23**, 30-38.

McGowan, R. S. (1994). "Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests", *Speech Comm.* 14, 19-48.

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*, Doctoral Thesis, The University of New South Wales, Australia.

Mokhtari, P. & Campbell, N. (2002a). "Perceptual validation of a voice quality parameter AQ automatically measured in acoustic islands of reliability", in *Proc. Meeting of the Acoust. Soc. of Japan*, Kanagawa Univ, Japan, Paper 1-P-25, 401-402.

Mokhtari, P. & Campbell, N. (2002b). Paper submitted to the International Conference on Spoken Language Processing, Denver, Colorado, USA.

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

Ohman, S. (1967). "Numerical model of coarticulation", *J. Acoust. Soc. Am.* **41**, 310-320.

Peterson, G. E. & Shoup, J. E. (1966). "A physiological theory of phonetics", *J. Speech Hear. Res.* **9**, 5-67.

Schoentgen, J. & Ciocea, S. (1997). "Kinematic formant-to-area mapping", *Speech Comm.* **21**, 227-244.

Schroeder, M. R. (1967). "Determination of the geometry of the human vocal tract by acoustic measurements," *J. Acoust. Soc. Am.* **41**, 1002-1010.

Snack software package, http://www.speech.kth.se/snack/.

Sundberg, J. (1987). *The science of the singing voice*, Northern Illinois University Press, Dekalb, Illinois.

Yehia, H. & Itakura, F. (1996). "A method to combine acoustic and morphological constraints in the speech production inverse problem", *Speech Comm.* **18**, 151-174.