PAPER *Special Issue on Speech Information Processing*

Automatic Measurement of Pressed/Breathy Phonation at Acoustic Centres of Reliability in Continuous Speech

Parham MOKHTARI[†*], ***Nonmember***, **Nick CAMPBELL**[†], ***Nonmember***

Summary

With the aim of enabling concatenative synthesis of expressive speech, we herein report progress towards developing robust and automatic algorithms for paralinguistic annotation of very large recorded-speech corpora. In particular, we describe 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. We then report some evaluations of a specific voice-quality parameter known as the glottal Amplitude Quotient (AQ), which was proposed in [2, 6] and is here measured automatically at centres of reliability in continuous speech. Analyses of a large, single-speaker corpus of emotional speech first validate the perceptual importance of the AQ parameter in quantifying the mode of phonation along the pressed-modal-breathy continuum, then reveal some of its phonetic, prosodic, and paralinguistic dependencies.

Key words:

voice-quality, breathiness, emotions, paralinguistic annotation

1. Introduction

Key elements of a concatenative speech synthesis system include the creation and appropriate annotation of spoken language data. As part of that effort, and in the framework of the Japan Science and Technology (JST) Corporation's project on Expressive Speech Processing (ESP), we are gathering very large amounts of natural speech recorded by subjects in everyday situations [5]. While it is thus intended to capture statistically significant samples of a wide range of emotional attitudes and speaking styles, practical considerations on the creation of such huge corpora also demand acoustic speech signal processing algorithms which are sufficiently robust, and which require minimal amounts of manual intervention. Such robust and automated algorithms are required first offline, to appropriately segment, parameterise, and annotate the speech data; then online, when it is required to judiciously select and concatenate speech units in order to synthesise a desired utterance in a desired speaking style or emotion.

To gain the flexibility of synthesising speech in a variety of speaking-styles, phonetic and prosodic labels must be augmented with an additional layer of paralinguistic tags which serve to differentiate various voice-qualities. Although the ultimate test of the efficacy of such tags is perhaps an auditory-perceptual evaluation of the synthesised speech, consistent with Laver's [11] descriptive framework we believe that the annotation process itself must respect the fact that the origin of variability in voice-quality lies in the speech production domain. Our approach is therefore to map the measured acoustics of speech to physiologically-related parameters of the vocal-tract area-function and the glottal-source waveform; to model paralinguistic variabilities in those parameters; and then to test their perceptual relevance and perhaps to refine them, by auditory evaluations.

Although the speech inverse problem of mapping from acoustics to articulatory/phonatory parameters is far from resolved, the most direct methods reported in the literature over the past forty years or so have relied largely on the formants (or acoustic resonances of the vocal-tract). However, while the formants are arguably the acoustic parameters most closely interpretable in articulatory terms, they are admittedly very difficult to measure reliably and robustly throughout continuous speech, even in voiced segments where formants may be expected – hence the need for supervision and manual correction when reliable measurements are required; and hence the tendency in studies using large amounts of speech, to relinquish the formants in favour of more easily measured acoustic parameters such as the cepstrum. One promising approach, inspired by Lea's [12] concept of "islands of phonetic reliability", is to access the formants wherever they may be most reliably measured in the speech stream.

In this paper we report progress towards developing robust algorithms for automatic measurement of physiologically-motivated voice-quality parameters in natural, emotional speech. In particular, we focus on the following two aspects: (i) automatic location of acoustic centres of reliability, and (ii) automatic estimation of a phonatory voice-quality parameter at those reliable centres. In section 2 we motivate and describe methods of locating centres of reliability in a recorded corpus of emotional speech. In section 3 we describe a method of estimating the glottal Amplitude Quotient (AQ) at those reliable centres; we then report on the phonetic-, prosodic-,

Manuscript received July 1, 2002.
Manuscript revised September 20, 2002.
[†]The authors are with the JST/CREST-ESP Project, at ATR Human Information Science Laboratories, 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288, Japan.
[*]E-mail: parham@atr.co.jp

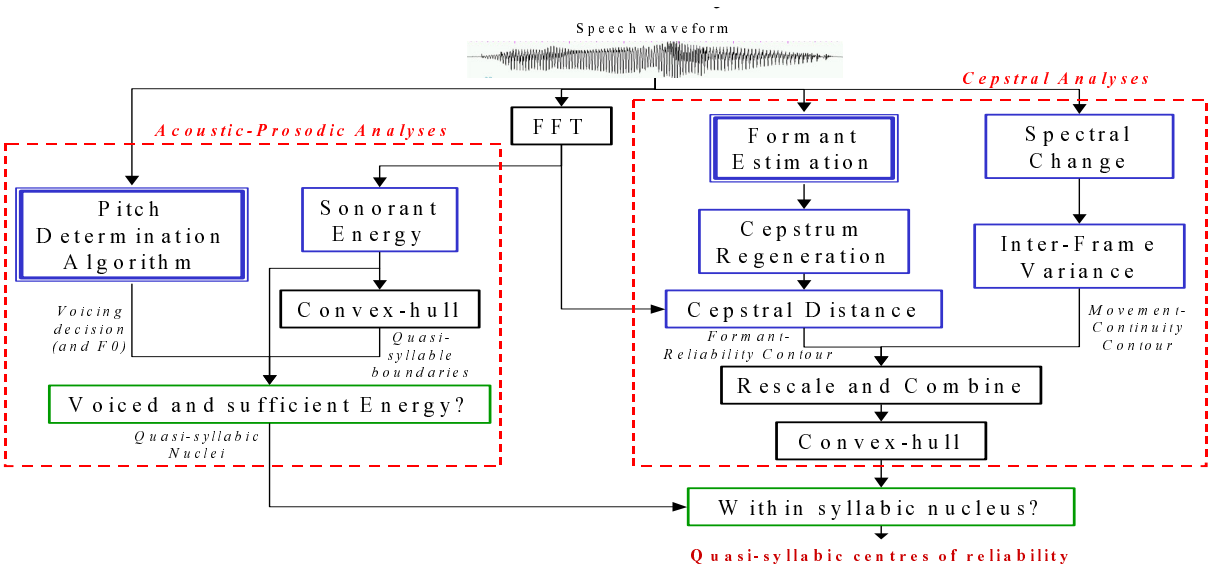


Fig. 1 Flow-chart of an algorithm to locate reliable centres in continuous speech (see section 2 for details).

and emotion-induced variabilities in that parameter, and show results of perceptual tests which confirm its salience as an acoustic correlate of perceived phonation quality along the pressed-breathy continuum. In section 4 we offer a summarising perspective and point to some ongoing and future research.

2. Centres of Reliability

In this section we first provide some computational definitions, then describe our algorithm for locating centres of reliability in the acoustic speech stream.

2.1 Computational Definitions

In the research context introduced above and in agreement with Öhman's thesis [18], we view the speech stream in its primitive form as a continuum of vocoids and contoids. While contoids can be regarded as articulatorily constrictive perturbations in the speech stream, vocoids are the more likely candidates for syllabic nuclei, generally exhibiting relatively stable spectral continuity and relatively high levels of acoustic energy. It is within just such portions of the acoustic speech stream – the more likely candidates for what Lea et al. [12, 14] have termed "islands of reliability" – that the perceptually and articulatorily salient, formant parameters can be measured with the greatest reliability; and by extension, these portions 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 continuity; and (iii) yield good initial estimates of the formants. Specifying these criteria in acoustic terms, our rationale prescribes: (i) a vocoid or fully-voiced region with relatively high sonorant-energy; (ii) either a spectral steady-state or a region with relatively smooth spectral-change; and (iii) a region where initial formant estimates provide a close match with respect to the original acoustics, e.g. in a spectral-matching sense. In the next section we describe in detail our algorithm which follows directly from these computational definitions.

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 with *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.

At the heart of the acoustic-prosodic strand is the quasi-syllabic segmentation method of Mermelstein [15], wherein the so-called convex-hull algorithm detects 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 [60, 3000] Hz which, after some speaker-specific tuning, is intended to encompass the sonorant range from the fundamental frequency F0 up to about the third formant F3, while largely excluding the higher-frequency energies associated 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 (according to a waveform-correlation threshold used in conjuction with a pitch-determination

algorithm [8] based on sub-harmonic summation) 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 [13], where its benefits for acoustic-prosodic analysis were demonstrated.

Meanwhile in the strand dominated by cepstral analyses, the articulatory-phonetic concepts of steady-state and continuous-movement [19] 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 computing a measure of spectral change, as afforded simply by the delta-cepstrum; the local (dis)continuity in spectral change is then quantified using a quefrency-weighted cepstral distance measure [23] to compute an inter-frame variance [16] 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 proposed by Broad & Clermont [4] and later pursued in [3] and [9]. Particularly in the speaker-dependent case, the linear mapping from LP-cepstrum to formants is remarkably robust [4], in the sense that while the estimated formants are not highly accurate, nor are they grossly incorrect (as can often occur in conventional formant estimation methods based on spectral-matching, where formants can be missed or assigned to the wrong spectral peak, even when dynamic constraints are used). It is precisely this type of robustness which is desirable for unsupervised analyses of large speech corpora, and which we here find indispensable. Those estimated formants are then used to compute simplified LP-cepstra, which are compared frame by frame with the corresponding FFT-cepstra computed from the original speech waveform. That comparison using a (quefrency-weighted) cepstral distance measure yields a contour of initial formant-(un)reliability, returning for this purpose to the 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 (cf. panel "*a*" in Fig. 2) and the contour of initial-formant reliability (cf. panel "*b*" in Fig. 2) – are each linearly rescaled to the range [0,1] then 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 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 subsequent analyses of voice quality, one example of which is described later in section 3.

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

	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

2.3 Phonetic Distribution

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 [10]. Each of three, read stories were designed to naturally evoke the emotions Anger, Joy, and Sadness; 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 coincided 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 segmentation and labelling information, it is interesting 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 centres detected in this category, the nasals are the most common, followed by the liquids and semi-vowels. As for the five vowels, the majority coinciding with reliable centres are /a/ and /o/ with around 6000 of each, followed by /e/ and /i/ with around 3000 of each, and finally /u/ represented by just under 2000 reliable centres. While the great majority (93.2%) of reliable centres thus coincide with the five vowels, it is important to note that our algorithm was not intended as a vowel-spotter, and that the centres which fall

into the "other" category are therefore not errors as such; rather, each detected centre is to be interpreted as a part of the speech stream where, regardless of the phonetic identity, acoustic (particularly formant) measurement reliability is relatively higher than other parts.

As shown in the example snapshot 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; in the 2.7sec interval shown, about 10 quasi-syllables and 9 reliable centres are located. Evidence of the fact that in this experiment we have set parameters of the algorithm to conservative values, can be found for example in the 2nd and the 5th quasi-syllables where, despite the relatively high plateau of sonorant energy, the pitch-detection algorithm (and its subsequent 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 2nd syllable. Also, the detected 7th and 8th nuclei appear to have durations longer than a human linguist might decide, the former spanning a portion of the utterance labelled /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 be equal to or 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, 9th detected syllable, the devoiced vowel in "shik" is nevertheless syllabified by virtue of the plateau in sonorant energy around the high F2 and F3 region.

3. Mode of Glottal Phonation

In this section we first motivate and describe our methods for measuring the glottal Amplitude Quotient (AQ); we then remark on its global distribution, provide a perceptual evaluation, then examine some phonetic-, prosodic-, and emotion-related variabilities in that parameter measured automatically in continuous speech.

3.1 Motivations

As described in section 2, the purpose of our algorithm to automatically locate *centres of reliability* in continuous speech is to achieve robust and reliable measurements of various types of *voice-quality*, thanks to the greater reliability of the formants measured around those centres. One aspect of voice-quality which is highly relevant to the expressivity of the speaker and therefore to the synthesis of emotional speech, is the state of the glottal voice-source along the continuum from a *pressed*, through *modal*, to a *breathy* phonation. While a great deal of research has been reported on the physiological and acoustic cues which

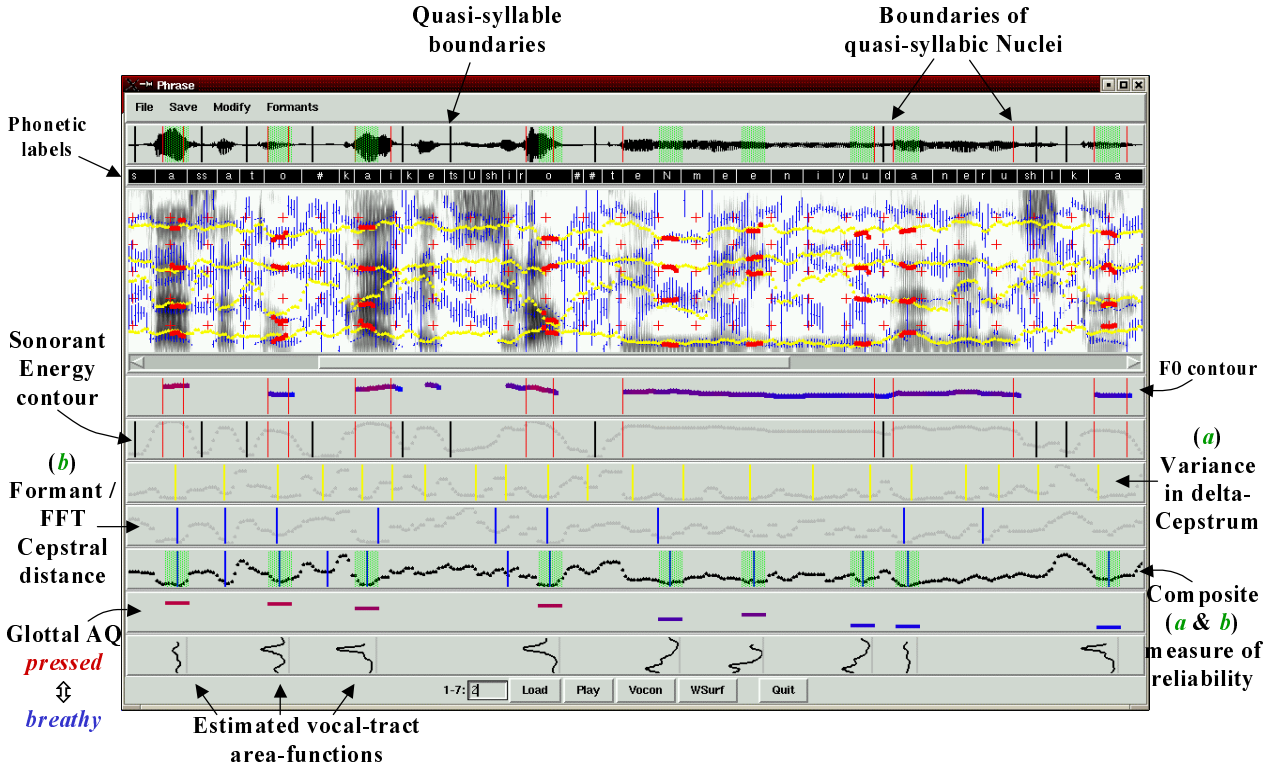


Fig. 2 Snapshot of our graphical user interface, implemented in Tcl/Tk and also using some display utilities provided in the Snack package [21]. See text for a description of each panel.

differentiate breathy from pressed voice (e.g., for an authoritative overview, cf. Sundberg [22]), the majority of such studies are limited to speech (and singing) recorded during sustained phonation of steady-state vowels. It indeed remains a challenge to robustly quantify the degree of pressedness or breathiness, automatically from acoustic measurements in large amounts of spontaneous speech.

While there are various measures which approximate voice-source properties in the spectral domain [20] [7], the glottal-flow waveform provides the most direct estimates. In particular, Fant et al. [6] and Alku & Vilkman [2] have proposed an Amplitude Quotient (AQ) which is the peak-to-peak amplitude of the glottal-flow waveform divided by the minimum amplitude of the flow derivative. One advantage of the AQ is that, owing to its relative independence of the sound pressure level which is related mainly to the denominator of the quotient [22], it quantifies mainly phonation *quality*. Another advantage is that it is an amplitude-domain parameter and should therefore be immune to the sources of error in measuring time-domain features of the estimated and often stylised glottal waveform. Moreover, both the numerator and the denominator of AQ "are rather insensitive to errors in inverse filter tuning" [6, p.1453]. Finally, for all of 4 male and 4 female speakers producing the sustained vowel "a" with a range of phonation types, "the value of AQ decreased monotonically when phonation was changed from breathy to pressed" [2, p.136].

3.2 Measuring the Glottal Amplitude-Quotient (AQ)

For every group of 5 consecutive analysis-frames around each reliable centre, an AQ value may be computed from an estimate of the glottal waveform. However, the reliability of such an estimate itself depends on the success with which the effects of the vocal-tract resonances can be eliminated from the corresponding 64msec of speech. While the formants at those 5 frames are by definition already well estimated, we further optimise them by small perturbations that minimise the distance between a regenerated LP-cepstrum and the FFT-cepstrum. After high-pass filtering of the original speech at 70Hz to eliminate rumble, followed by adaptive low-pass filtering to reduce information above the 4th formant, the optimised formants are used to construct a time-varying inverse-filter which cancels the effects of the first 4 formants. The result – an estimate of the glottal waveform derivative – is simply integrated to estimate the glottal-flow waveform. Finally, AQ is computed as the ratio of the largest peak-to-peak amplitude of that waveform and the largest amplitude of the cycle-to-cycle minima of the glottal-flow derivative.

3.3 Global Distribution of AQ

Figure 3 shows a histogram of all the AQ values automatically measured at the 22175 centres of reliability reported earlier. The values range from 2.2 (a more pressed phonation) to 24.7 (a more breathy phonation), with a mean of 8.8 (a prototypical or modal phonation). While the distribution is unimodal, there does appear to be a more gradual descent towards the higher (more breathy) end. Indeed, although not shown here, it is interesting to note that a more closely Gaussian distribution was obtained simply by transforming AQ onto a logarithmic scale; however, the physical and perceptual implications of a logarithmic AQ scale remain to be clarified.

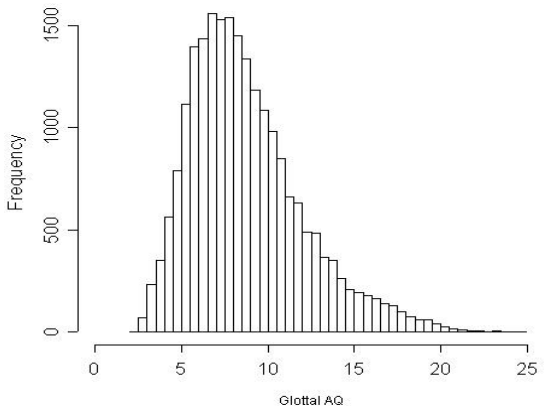


Fig. 3 Distribution of AQ measured in the 22,175 reliable centres automatically located in our speech database.

3.4 Auditory-Perceptual Evaluation of AQ

Combining the results of groups of reliable-centres lying within the same quasi-syllabic nuclei, and disregarding centres at which the composite measure of unreliability is above an empirically determined threshold, the number of nuclei that can potentially be used as auditory stimuli to test the perceptual efficacy of AQ was reduced to just over 15,000. Statistics computed over this dataset allowed the selection of 60 stimuli to be used for perceptual evaluation. In particular, for each of the 3 emotions, 5 nuclei were chosen whose reliable-centres had a value of AQ in either of the following 4 categories: extremely low; extremely high; around the mean minus one standard-deviation of that emotion's AQ distribution; around the mean plus one standard-deviation.

The durations of the 60 quasi-syllabic nuclei thus selected, ranged from 32msec to 560msec, with a mean of 171msec. Eleven normal-hearing subjects participated in an auditory evaluation of these short stimuli, listening to each one as many times as required over high-quality headphones in a quiet office environment, and rating each on two separate, 7-point scales which were explained simply as "perceived breathiness" and "perceived loudness", respectively. The ratings of each subject were then linearly rescaled onto the range [0,1], and these normalised scores were averaged across all 11 subjects to obtain a single, representative value of breathiness and of loudness for each of the 60 stimuli.

Figure 4 shows a scatter-plot comparing perceived breathiness with the acoustically-measured values of AQ.

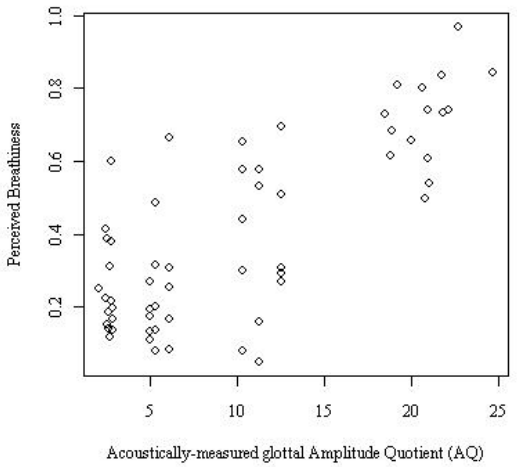


Fig. 4 Scatter-plot comparing the perceived degree of breathiness with the automatically-measured, glottal AQ parameter in all 60 auditory stimuli. (r = 0.77).

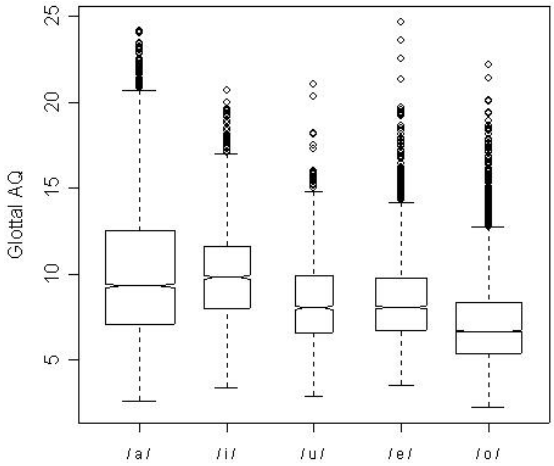


Fig. 5 Vowel dependence of AQ. Each box is delimited vertically at the first quartile each side of the median shown by the centre-line, its width is proportional to the square-root of the number of samples, and whiskers delimit 1.5 times a quartile each side of the median.

The linear coefficient of correlation for these 60 pairs of values was found to be 0.77. While this correlation is not particularly high, it does characterise the strength of an obvious trend that as the measured AQ increases, so too does the perceived breathiness of the speech stimulus on average. A closer examination of some of the points which lie furthest from a line of best fit, revealed some potential causes of error: formant discontinuities across the 5 frames, owing to a lack of dynamic constraints in formant estimation; a higher degree of breathiness during a part of the syllabic nucleus not included in the 5 frames; strong influence of adjacent nasality on the vowel portion within the 5 frames. Furthermore, it is interesting to note in Figure 4 that there is a greater range of perceived breathiness for those stimuli with a mid-to-low AQ value, confirming intuition that it is a more difficult task to rate the breathiness of stimuli which are perhaps better characterised by either *modal* or *pressed* phonation.

The correlation similarly computed between the perceived loudness and the RMS-energy measured at the same reliable-centres, was found to be 0.83, thus confirming the strength of that relation despite not having used a more sophisticated, perceptually weighted measure of loudness. To make a link with the previous results, we were interested in testing whether the AQ values were at all correlated with the RMS-energy in the original speech waveform, or whether, as claimed in [2], AQ is largely independent of the signal energy or sound pressure level. The correlation between the acoustically-measured RMS and AQ values was found to be –0.70, indicating a weaker but inverse relation. However, the homologous correlation between the *perceived* loudness and breathiness ratings was found to be –0.78, suggesting that the inverse relation is not just an artefact in computing AQ. At least for these data, it seems that perceived breathiness (and measured AQ) is negatively related with perceived loudness (and measured RMS-energy); the cause-and-effect properties underlying this relation remain to be investigated.

3.5 Phonetic Dependence of AQ

The boxplot in Figure 5 compares the distribution of AQ across the five vowels, i.e., the phonetic categories for which the estimates of the formants, glottal waveforms and AQ values may be assumed to be the most reliable. While there is a large amount of overlap in the total range of values, there appears a trend with /a/ and /i/ having a higher distribution (more breathy) and /o/ having a lower distribution (more modal or pressed).

Considering aerodynamic principles of the vocal-tract, one might expect an inherently more breathy phonation in vowels that are produced with a greater degree of oral constriction and thus a reduced transglottal difference in air-pressure; conversely, one might expect an inherently less breathy (or more modal) phonation for less-constricted (or more neutral) vowels, where a higher transglottal air-pressure promotes greater efficiency in glottal vibrations. Although the data in Figure 5 do not immediately support these hypotheses, we did find that the distribution of AQ was indeed slightly higher (more breathy) for all the vowel samples lying around the periphery of the F1-F2 plane (mean AQ of 9) than for the remaining, more neutral vowel samples whose F1 and F2 lay inside boundaries defined by the middle two quartiles in each of those dimensions (mean AQ of 8).

Another factor which might be expected to have an influence on AQ comparable in magnitude with that of vowel-inherent phonation, is the relative position of the segment within each phrase. To investigate this question, we identified those reliable centres whose phonetic label either followed or preceded a silence-label (to within two labels). This crude analysis yielded 3160 phrase-initial and 2797 phrase-final candidates. Interestingly, the

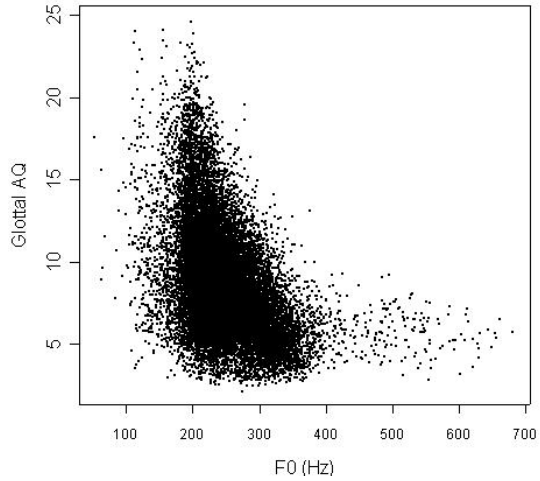


Fig. 6 Distribution of the AQ parameter in all the 20,674 vowel samples, as a function of the mean F0 measured around the corresponding centres of reliability.

phrase-final segments were found to have a higher distribution of AQ (mean 10.5), and therefore presumably a breathier voice-quality on average, than the phrase-initial segments (mean 9.1).

3.6 Prosodic Dependence of AQ

Once again considering only those 20674 samples of more reliable data which were found to coincide with one of the five vowels, Figure 6 shows the dependence of AQ on the fundamental frequency of voicing. An interesting trend that emerges from the scatter-plot is that vowel segments with higher values of F0 are more likely to have a low AQ, i.e., high F0 is more likely accompanied by either modal or pressed phonation. Conversely, the tapered distribution towards the upper-left corner of the plot reveals that vowel segments having higher values of AQ (a breathier voice) are more likely to be produced with a lower F0. Indeed: "High pitched breathy voices seem rare" [11, p.133].

The approximately inverse relation between AQ and F0 portrayed in Figure 6, in fact provides empirical support to the recent proposal by Alku et al. [1] to divide AQ by the local pitch period, or equivalently, to multiply it by the local F0. The resulting Normalised Amplitude Quotient (NAQ) has been shown to be closely related with the well-known glottal Closing Quotient [1], and should therefore be an even better parameter than AQ in regard to quantifying phonation type independently of F0. However, as the interpretations of all our results were found to be essentially the same when considering the distributions of either AQ or NAQ, pending further detailed investigations we shall continue in this paper to report our results using the (unnormalised) AQ parameter.

3.7 Paralinguistic Dependence of AQ

Our expressive speech database allows us next to compare the distributions of AQ measured in each of the three emotion categories Anger, Joy, and Sadness. Before considering the data, it is instructive to note two points which may undermine or weaken any direct relationship between emotions and AQ. First, each of the three datasets inevitably contains expressive qualities other than the main emotion intended by the respective story; at the very least, they each would contain many examples of emotionally *neutral* speech [10]. Second, even if all the utterances in each dataset were to convey strongly the intended emotion, it would be naive to assume that a speaker will consistently employ categorically distinct modes of phonation to convey those three emotions.

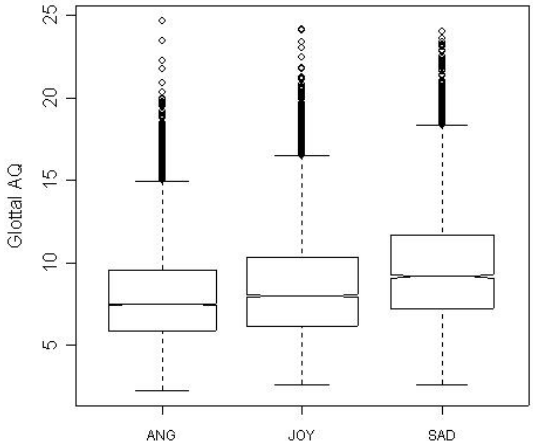


Fig. 7 Dependence of AQ on intended emotion. Boxplot details are the same as those for Figure 5.

Despite these reservations, the boxplot in Figure 7 reveals an interesting trend: the distribution of AQ tends to higher values in the order Anger (mean 8.0), Joy (mean 8.6) and Sadness (mean 9.8), with a particularly marked increase in the upper-quartile range. As expected, these data do not show a categorical distinction that might be directly exploited, e.g. in an emotion-recognition system. However, they do portray a clear trend that our speaker's Sad speech is more likely to be biased towards breathiness, and that her Angry speech is by contrast more likely to contain either modal or pressed phonation. While these results are intuitively appealing, it will also be interesting in future work to compare the emotion-related use of phonation quality across different speakers.

4. Summary and Outlook

In this paper we described a method of combining robust acoustic-prosodic and cepstral analyses to first locate centres of reliability in continuous speech, where measured formants yield more reliable estimates of physiologically-related voice-quality parameters. We then described a method of estimating the glottal AQ parameter [2, 6] at those reliable centres, and reported an auditory evaluation which showed its effectiveness in quantifying the perceived degree of breathiness in glottal phonation. Finally, we examined the distribution of AQ as a function of phonetic, prosodic, and emotion-related variabilities.

Our unsupervised algorithms which first locate centres of reliability in speech, have enabled a more extensive empirical investigation of the glottal AQ parameter than has hitherto been reported.

Ongoing research is aimed primarily at incorporating such automatic measures of voice-quality in the processes of database annotation and unit selection in concatenative speech synthesis. Towards that aim, we are developing methods of robustly quantifying not only phonatory but also supralaryngeal voice-qualities or articulatory settings [11] using estimated vocal-tract area-functions (cf. bottom panel in Figure 2) and a tripartite model of variability [16, 17]; and we are extending our investigations to annotate larger, more diverse and more spontaneous databases of speech recorded by a greater number of speakers. Bearing in mind the complexities of human emotions manifested in speech, we are particularly interested in unfolding the intricate relationships amongst physiologically-related voice-quality parameters which shape, more so than determine, a speaker's expressions and speaking styles. It is hoped that a better understanding of these issues will aid in the development of a speech synthesiser able to more faithfully reproduce various expressive speaking styles.

Acknowledgements

This work is supported by the Japan Science and Technology (JST) Corporation under CREST Project 131.

References

[1] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow", J. Acoust. Soc. Am., vol.112, no.2, pp.701-710, 2002.

[2] P. Alku and E. Vilkman, "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering", Speech Comm., vol.18, no.2, pp.131-138, 1996.

[3] A. Bayya and H. Hermansky, "Towards feature-based speech metric", in Proc. IEEE Int. Conf. on Acoust., Speech, and Sig. Process., pp.781-784, 1990.

[4] D. J. Broad and F. Clermont, "Formant estimation by linear transformation of the LPC cepstrum", J. Acoust. Soc. Am., vol.86, no.5, pp.2013-2017, 1989.

[5] N. Campbell, "Recording techniques for capturing natural every-day speech", in Proc. 3rd Int. Conf. on Lang. Resourc. and Eval., Las Palmas, Spain, pp.2029-2032, 2002.

[6] G Fant, A. Kruckenberg, J. Liljencrants and M. Bavegard, "Voice source parameters in continuous speech. Transformation of LF-parameters", in Proc. 3rd Int. Conf. On Spoken Lang. Process., pp.1451-1454, 1994.

[7] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates", J. Acoust. Soc. Am., vol.101, no.1, pp.466-481, 1997.

[8] D. Hermes, "Measurement of pitch by subharmonic summation", J. Acoust. Soc. Am., vol.83, no.1, pp.257-264, 1988.

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

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

[11] J. Laver, The phonetic description of voice quality, CPU, Cambridge, 1980.

[12] W. A. Lea, "Prosodic aids to speech recognition", in W. A. Lea, ed., Trends in speech recognition, Prentice-Hall, New Jersey, pp.166-205, 1980.

[13] W. A. Lea and F. Clermont, "Algorithms for acoustic prosodic analysis", in Proc. IEEE Int. Conf. on Acoust., Speech, and Sig. Process., pp.42.7.1-42.7.4, 1984.

[14] W. A. Lea, M. F. Medress, and T. E. Skinner, "A prosodically guided speech understanding strategy", IEEE Trans. on Acoust., Speech, and Sig. Process., vol.23, pp.30-38, 1975.

[15] P. Mermelstein, "Automatic segmentation of speech into syllabic units", J. Acoust. Soc. Am., vol.58, no.4, pp.880-883, 1975.

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

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

[18] S. Öhman, "Numerical model of coarticulation", J. Acoust. Soc. Am., vol.41, pp.310-320, 1967.

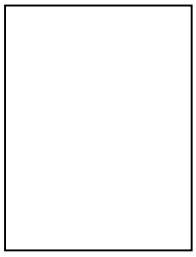
[19] G. E. Peterson and J. E. Shoup, "A physiological theory of phonetics", J. Speech Hear. Res., vol.9, pp.5-67, 1966.

[20] A. Sluijter, Phonetic correlates of stress and accent, (The Hague: HIL), Holland, 1995.

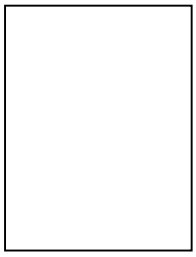
[21] Snack software package, http://www.speech.kth.se/snack/

[22] J. Sundberg, The science of the singing voice, Northern Illinois University Press, Dekalb, Illinois, 1987.

[23] B. Yegnanarayana and D. R. Reddy, "A distance measure based on the derivative of linear prediction phase spectrum", in Proc. IEEE Int. Conf. on Acoust., Speech, and Sig. Process., pp.744-747, 1979.



Parham Mokhtari received the B.E. degree in Electronics and Communications Engineering from the Univ. of Canberra (Australia) in 1993, and the PhD in Computer Science from UNSW in Canberra in 1998. From July 1998 to February 1999 he was an INRIA Postdoctoral Fellow at Loria, Nancy (France); he then held an STA Postdoctoral Fellowship at ETL in Tsukuba (Japan) for two years; and since April 2002 he is a research scientist in the JST/CREST ESP Project at the Advanced Telecommunications Research Institute International (ATR) in Kyoto, Japan. His research interests include acoustic and articulatory characterisations of phonetic variability, speaker individuality, and human emotions in speech.



Nick Campbell is currently a Project Leader at the ATR Human Information Science Laboratories, and Research Director for the JST/CREST Expressive Speech Processing project. He received his PhD in Experimental Psychology from the University of Sussex in the U.K. He was invited as a Research Fellow at the IBM UK Scientific Centre, where he developed algorithms for speech synthesis, and was an invited researcher at the AT&T Bell Laboratories. He served as Senior Linguist at the Edinburgh University Centre for Speech Technology Research before joining ATR in 1990. His research interests include large speech databases, concatenative speech synthesis, and prosodic information modelling. He spends his spare time working with postgraduate students as Visiting Professor at NAIST and Kobe University.