Some articulatory correlates of emotion variability in speech: a preliminary study on spoken Japanese vowels

Parham Mokhtari[†], *Akemi Iida*[‡] *and Nick Campbell*[†]

Japan Science and Technology Corporation (JST) CREST Project on Expressive Speech Processing (ESP)

[†]Advanced Telecommunications Research Institute Int. (ATR), 2-2 Hikaridai, Seika-cho, Kyoto 619-0288, Japan.

[‡]Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa, Kanagawa 252-8520, Japan.

parham@isd.atr.co.jp akeiida@sfc.keio.ac.jp nick@slt.atr.co.jp

ABSTRACT

An acoustic-articulatory model previously proposed to account for inter-speaker variability (Mokhtari et al., 2000), is here reinterpreted to model *intra-speaker, emotion-induced variability*. A decomposition of that variability in terms of the supralaryngeal vocal-tract's *longitudinal setting*, *latitudinal setting* (Laver, 1980), and vowel-specific articulatory *strategy*, affords a more complete descriptive framework, which is here applied to spoken vowel data recorded by an adult female speaker of Japanese, in the three emotional states Anger, Joy, and Sadness. We aim thereby to provide an articulatory-segmental labelling of source-unit databases, for more judicious selection of speaking-style in concatenative speech synthesis.

1. INTRODUCTION

As the intelligibility of computer-synthesised speech has for several years attained a reasonably acceptable level, research is increasingly directed at improving its naturalness. Advances over the last decade or so in particular have resulted in the growing popularity of so-called concatenative methods of text-to-speech synthesis (e.g., Campbell and Black, 1996), where individual segments of the speech stream are appropriately selected from a large database of natural, pre-recorded speech. However, while such re-sequencing of recorded speech affords more natural-sounding synthesis faithful to the characteristics of the recorded speaker, there remain challenges in the design, construction, and labelling of such databases (Campbell, 2000) in order to enable not only an unlimited range of linguistic recombinations, but also a sufficiently wide coverage of speaking styles as conveyed for example by para-linguistic and extra-linguistic information.

One of the outstanding challenges in that arena is the synthesis of speech incorporating human vocal emotions. Particularly in the application of synthesis as a communication aid for the orally handicapped, it has become clear that the speaking-style coverage of a conventionally-constructed, read-text database is insufficient for many of the user's daily communication requirements (Iida et al., 1998), and that a highly desirable improvement would be to gain flexible control of the expressive component or speaking-style. In that regard, the frequently cited and excellent review of the literature on human vocal emotions by Murray and Arnott (1993) is one important testament to the tendency to consider predominantly the prosodic, rather than the segmental strand of speech analysis in search of the acoustic manifestations of emotion variability. In particular, much of the relevant literature is focussed on the acoustic-prosodic parameters fundamental frequency (F0) of voicing, segment duration, energy and, to a lesser extent, laryngeal voice-quality; and indeed, the literature has repeatedly confirmed the significance of those prosodic parameters to perceived emotion. However, while relatively much less attention has been paid to the segmental correlates of emotion variability, a number of studies have already pointed out the importance of the supra-laryngeal vocal-tract (VT) in conveying emotional expression in speech.

One early example is the study by Fónagy (1976), who interpreted articulatory movements as seen on midsagittal X-ray films of two speakers uttering six Hungarian phrases with various acted emotions. Fónagy's gestural approach led him to describe whole movements of the tongue in such terms as "rising slowly, feebly" in sad speech, with "rapidity and a certain tension" in joyful speech, and to "suddenly lance forward like an arrow" in menacing speech. In a similar vein, Murray and Arnott (1993, Table I, p.1106) summarised the effects of vocal emotion on articulation with the following adjectives: "normal" in happiness and in disgust, "tense" in anger, "slurring" in sadness, and "precise" in fear.

More quantitative analyses of direct articulatory measurements obtained with an EMMA system were reported recently by Erickson et al. (2000) for American English, and by Maekawa and Kagomiya (2000) for Japanese. In the former study, articulatory and acoustic (formant) measurements were taken at the point of maximum jaw opening in the vowel portion of the first word and in the first syllable of the second word in the sentence "That's wonderful" repeated with several different intonation contours and simulated emotions; the authors observed greater jaw lowering in anger, tongue raising in suspicion, and tongue lowering

in admiration. In the latter study on Japanese, measurements were taken at the point of maximum jaw opening in the vowel portions of four isolated words and one short phrase; scatter-plots of the first two formants and of the midsagittal position of a coil placed on the tongue-dorsum both revealed that the speaker's tongue took "a relatively more frontal position in Suspicion than in Admiration", at least in the vowels /a/, /e/, and /i/ analysed; in addition, it was noted that the distance between the upper and lower lips was generally greater in suspicion than in admiration.

As direct articulatory measurements require specialist equipment not always readily accessible, a number of studies have also relied on articulatory interpretations of acoustic data. For example, Kienast and Sendlmeier (2000) measured the first two formants at the vowel mid-points of three sentences recorded with various emotions by six speakers of German; they observed a centralisation or reduction of the vowel formant space in those sentences expressing fear, sadness and boredom, and an expansion towards more extreme formant values in those sentences expressing anger. Perceptual validations of the expressive effect of vowel formant reduction have in turn been reported by Burkhardt and Sendlmeier (2000) using a formant synthesiser, and by Rank and Pirker (1998) using a concatenative synthesiser.

However, one limitation common to all the studies cited above, regardless of the availability of direct articulatory measurements, is the apparent lack of a general, acoustic-articulatory framework within which emotion variability may be described more completely. A similar observation in regard to the problem of inter-speaker variability led Mokhtari et al. (2000) to propose an acoustic-articulatory model which decomposes that variability in terms of its three basic, supra-laryngeal sources. Indeed, using that model, the inter-speaker variability in vocal-tract area-functions estimated from the first four, carefully-measured formants of vowel steady-states, was decomposed and thus explained in terms of vocal-tract structure, overall articulatory setting, and phoneme-specific articulatory strategy.

In this paper, we propose to apply that methodology to investigate articulatory correlates of *intra-speaker, emotion variability*. As described in the following sections, we first construct a carefully-measured corpus of vowel formant-patterns from a database of emotional speech recorded by a speaker of Japanese. Those formant data are then used to estimate VT area-functions, which in turn form the basis of our acoustic-articulatory investigation of emotion variability in speech.

2. SPEECH DATA & COMPUTATIONAL METHODS

2.1. Speech Materials & Formant Estimation

The speech data (Iida et al., 1998) comprise three stories read by an adult, female, native speaker of Japanese. The content of each story was designed to naturally evoke the emotions Anger, Joy and Sadness, respectively. While there was no explicit control over the phonetic composition of the three texts, each story contains more than 400 sentences, or more than 30,000 phonemes, which is sufficient to build a source-unit database for the CHATR concatenative text-to-speech synthesiser [3]; and indeed, perceptual evaluation of semantically neutral texts thus resynthesised from each recorded corpus yielded significant identification of the intended emotion (Iida et al., 1998).

With the aim of finding the longest-duration and therefore presumably the most steady-state (or least coarticulated) examples of each vowel, the manually-checked phonetic-segmentation files of each recorded database were automatically queried in search of all instances of consecutive (repeated) vowels VV, VVV, and VVVV (where V represents either of the 5 Japanese vowels /i, e, a, o, u/). All such segments for each vowel and each emotion, were then ordered from the longest- to the shortest-duration, and informal listening tests were conducted (by the first author, who is familiar with but not a native speaker of Japanese) to identify the 15 longest-duration segments which also carry auditory-perceptual characteristics of the intended emotion, whether in isolation or, more often, in sentential context.

A custom-developed interactive software (with a graphical user interface written in Tcl/Tk, and speech input/output and spectrogram-display functions implemented using the Snack [15] package) was then used to measure the first four formants of each vowel steady-state. The 5 most steady-state, consecutive frames were identified within each vocalic nucleus, with the aid of a cepstral measure of inter-frame variance (Mokhtari, 1998). The first four formants were then selected from amongst the poles of a selective-linear-prediction (SLP) analysis, on the basis of (i) narrow bandwidth, (ii) per-vowel limits on expected formant ranges for an adult, female speaker, (iii) inter-frame continuity (judged visually by inspection of the poles superimposed on the segment's spectrogram), and (iv) inter-segment consistency per emotion category (judged visually by inspection of a concatenated formant chart showing all four measured formants of all 5 steady-state frames of all 15 segments for the given vowel and emotion). The software tool was designed to secure as reliable formants as possible, by allowing interactive adjustments of the following analysis parameters: upper frequency limit of SLP

analysis (set to either 5 or 6 kHz; where the original speech waveforms were sampled at 16 kHz), order of SLP analysis (selected within the range 10 to 20), coefficient of preemphasis (default value 0.98; often reduced to as low as 0.0), frame advance (default value 8 msec; reduced to 5 msec in only one troublesome segment), and the position of the automatically-identified steady-state within the vocalic segment (often shifted manually to better approximate the given vowel target).

2.2. Estimation of Vocal-Tract Area-Functions

The first four formant frequencies and bandwidths thus measured (in each of the 5 frames × 15 segments × 5 vowels × 3 emotions = total 1125 formant-patterns) were then used to estimate the VT-length (VTL) and VT-shape, using a method of inversion based on the linear-prediction vocal-tract (LP-VT) model, together with a new, acoustically-relevant parameterisation of the VT-shape in terms of the first four, odd-indexed cosine and sine coefficients of the logarithmic area-function (Mokhtari, 1998; Mokhtari & Clermont, in preparation). The VTL was estimated using a criterion of minimum deviation from a neutral tube, and the area-scaling factor was determined by enforcing a zero-mean logarithmic area-function.

As noted previously (Mokhtari, 1998; Mokhtari et al., 2000), an important step in the inversion method is the pre-treatment of each formant bandwidth, by first averaging over the 5 frames, 15 segments, and 3 emotions for each vowel separately, then subtracting a so-called "closed-glottis" bandwidth determined by Hawks & Miller's (1995) empirically-defined equations (HM95). The notorious unreliability of bandwidth measurements is thereby acknowledged by retaining only their gross, phonetic variations, and thus retaining inter-emotion variability only in the formant frequencies; moreover, the resulting estimates of the "glottis-only" bandwidths can be considered to be more relevant in estimating VT area-functions with the LP-VT model, which has only a single (glottal) source of acoustic energy loss.

However, in following that methodology, it was found that the per-vowel means of our measurements of the *fourth* formant bandwidths were *lower* than the "closed-glottis" estimates provided by HM95 at the corresponding, mean formant frequencies, and that the prescribed subtraction would therefore incorrectly yield *negative*-valued bandwidths. Admittedly, the empirical data used by Hawks and Miller (1995) were limited to only the first *three* formants; hence, while our procedure had previously yielded satisfactory results for data which included the fourth formants of adult male speakers (Mokhtari, 1998; Mokhtari et al., 2000), the model mismatch seems here to be emphasised by the much higher, fourth formant *frequencies* of our female speaker. As a compromise, our speaker's per-vowel mean F4 values (overall mean of 4589 Hz) were scaled down to the overall mean F4 (3339 Hz) previously measured for 5 adult male speakers of Japanese (Mokhtari and Tanaka, 2000), and the fourth formant bandwidths were then corrected using the HM95 frequency-bandwidth curves for adult male speakers.

	B1	B2	B3	B4
/i/	33	36	16	27
/e/	104	83	10	78
/a/	177	77	84	228
/o/	124	151	169	107
/u/	97	107	52	37

Table 1. "Glottis-only" formant bandwidths per vowel, estimated by the methods described in section 2.2.

The resulting, per-vowel estimates of the "glottis-only" bandwidths of our female speaker are listed in Table 1. It is interesting to note the relatively low B3 of the front vowels /i/ and /e/, for which the third formant can indeed be expected to be affiliated with the VT cavity anterior to the palatal place of lingual constriction, and thus relatively less influenced by glottal losses. It is also interesting to note the relatively low B4 of the mid-high vowel /u/, which may similarly indicate a relative decoupling of glottal influence from the cavity most strongly affiliated with the fourth formant in that vowel produced by our speaker. In conclusion, the use of bandwidths is critical for enforcing uniqueness in the estimation of VT area-functions, and their pre-treatment by the method described above helps to secure more realistic area-functions by better matching the characteristics of the LP-VT model.

2.3. Articulatory model of emotion variability

The model used here to study emotion variability, is a reinterpretation of the inter-speaker articulatory model proposed by Mokhtari et al. (2000), to which we refer the reader interested in more elaborate computational details. Briefly, the parameterised (and thereby smoothed) VT area-functions estimated for the vowels in each of the three emotion categories, are subjected to a decomposition of the emotion-related variability in terms of: (i) VT *structure*, computationally defined as the overall mean VTL of the area-functions of each emotion; (ii) articulatory *setting*, computationally defined as the overall mean VT-shape of the area-functions of each emotion; and (iii) articulatory *strategy*, computationally defined as the per-vowel mean VTLs and VT-shapes of the area-functions of each emotion, after normalisation of the emotion

variability in *structure* and *setting*.

This tripartite descriptive framework can be used to study intra-speaker, emotion variability in the supralaryngeal VT, if it is thus reinterpreted: (i) as only a single speaker's data are used, the VT *structure* component of our model here no longer refers to VT anatomical differences, but rather to emotion-induced differences in *longitudinal setting* (Laver, 1980); (ii) as a result, the articulatory *setting* component of our model should more appropriately be defined as *latitudinal setting* (Laver, 1980); (iii) the third component of our model retains its original name, and is simply reinterpreted in terms of emotion-induced, residual differences in vowel-specific articulatory *strategy*. It is important to note that while Laver's (1980) detailed framework on the phonetic description of voice quality was a source of inspiration in defining the first two components of our model, our relatively more simple definitions of the longitudinal and latitudinal settings are a necessary compromise to achieve computational tractability while nevertheless aiming to take full advantage of the available articulatory data (i.e., estimated VT area-functions).

3. RESULTS

The results of our method of inversion and articulatory modelling of intra-speaker emotion variability are here presented by considering in turn, each of the three components outlined in the previous section.

(i) Our method of inversion yielded area-functions with an overall mean VTL of 13.1cm. Compared with an overall mean VTL of 18.2cm obtained using the same methodology on data of 5 adult male speakers of Japanese (Mokhtari et al., 2000), our adult female speaker therefore appears to have a shorter VTL by 28% on average. While this may seem a rather extreme reduction of the VT anatomical size, it is well to note that our female speaker's overall mean F3 and F4 (3239 Hz and 4589 Hz) are respectively 30% and 37% higher than those of the 5 adult male speakers cited above (2491 Hz and 3339 Hz); on the basis of the measured acoustic data, the reduction in estimated VTL therefore appears to be reasonable.

The overall mean vocal-tract lengths obtained for each emotion category separately, are as follows: 12.9cm for Anger, 13.1cm for Joy, and 13.4cm for Sadness. Thus it seems that our speaker's effective VTL during vowel production is about 2% shorter in angry speech and about 2% longer in sad speech, compared with her overall mean VTL. This apparently small variation can be compared with the inter-speaker variability in the mean VTLs of the 5 adult male speakers reported by Mokhtari et al. (2000), wherein one standard-deviation was found to be 0.6cm or about 3% of the overall mean VTL. Naturally, emotion-induced intra-speaker variability in *longitudinal setting* is indeed expected to be smaller than VT-structural differences between speakers.

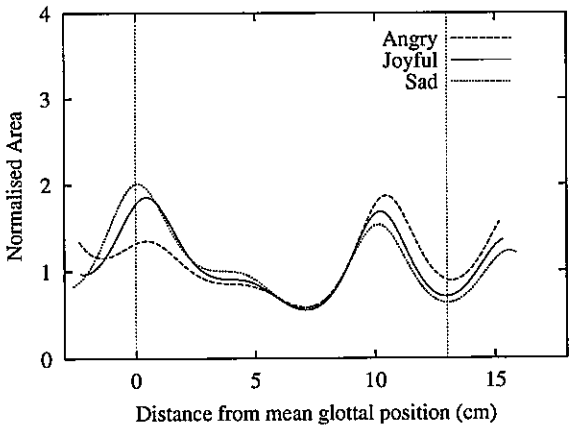


Figure 1. Emotion-related differences in articulatory *setting*. Shown superimposed are the overall mean (vowel-averaged) VT-shapes for each of the three emotion categories. The left and right, vertical lines mark the mean position of the glottis (at 0cm) and of the lips (at 13cm), respectively.

(ii) The results concerning supralaryngeal, *latitudinal setting*, are shown in Figure 1 by the vowel-averaged VT-shapes of the three emotion categories. Those superimposed area-functions suggest that our speaker's long-term articulatory settings in each of the three emotions converge to a nearly invariant, place of (quasi-) constriction at about 7-8cm from the mean glottal position. By contrast, emotion variability appears to be manifest in the back and in the front parts of the VT: our speaker's vowel production during angry speech seems to be characterised by a more constricted lower-pharynx, and by a greater degree of mouth opening – besides being intuitively satisfying, these results also agree with Erickson et al.'s (2000) observation of a greater degree of jaw opening in angry speech; and while the settings for joyful and sad speech appear to be relatively more similar to each other, the sad setting is the more extremely opposed to that of anger, displaying a more open lower-pharynx and a lesser degree of lip opening.

(iii) Emotion-related variability in articulatory *strategy* is defined in our model in terms of residual differences in VTL and VT-shape on a per-vowel basis. The strategy-related, mean VTLs listed in Table 2, show the largest inter-emotion variability in the vowels /e/ (shorter VTL in anger and longer VTL in sadness) and /a/ (shorter VTL in joy and longer VTL in anger). These variations may be explained in terms of vowel-specific articulatory behaviour in, for example, the amount of larynx raising/lowering or lip retraction/protrusion.

	Anger	Joy	Sadness
/i/	11.9	12.2	11.7
/e/	10.0	10.3	11.0
/a/	13.5	12.4	12.8
/o/	15.9	15.9	15.8
/u/	14.3	14.8	14.3

Table 2. *Strategy*-related, mean vocal-tract lengths (in cm) for each vowel and each emotion category.

We note in passing that the VTLs listed in Table 2 also display a reasonable range of phonetic variation, with the shortest mean VTL of 10.5cm for the mid-front vowel /e/ and the longest mean VTL of 15.9cm for the mid-back vowel /o/.

The strategy-related, mean VT-shapes of each emotion are shown for each vowel separately in the five panels of Figure 2. Emotion-induced variability appears in the main place of constriction in the vowels /e, u, a/, with joyful speech manifesting a consistently more fronted constriction location. There generally appears to be little variability in the degree of lip-opening, except for the open vowel /a/, which is most open in joyful speech. The front cavity between the lips and the main place of lingual constriction in the mid-vowels /e/ and /o/, appears to have the largest area in angry speech, and the smallest area in sad speech. By comparison with these obvious manifestations of variability, the regions of constriction in the vowels /i/ and /o/ exhibit very little, emotion-related variation, with only small differences in the degree of constriction for /o/.

4. DISCUSSION & APPLICATIONS

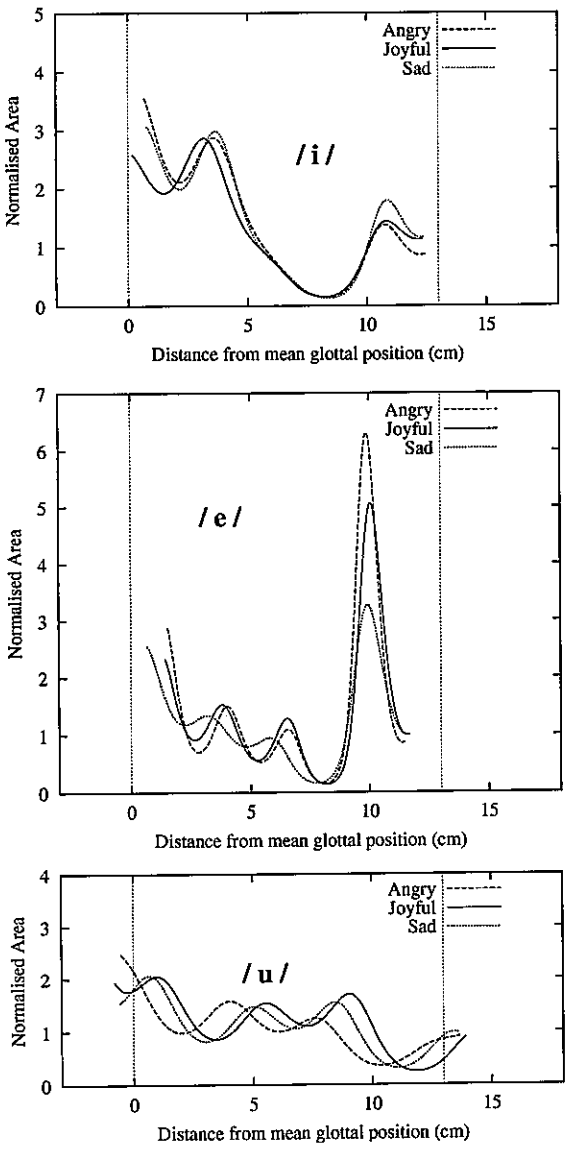
As intimated in the Introduction, our main motivation for the current research on segmental, articulatory correlates of emotion variability in speech, is to expand the utility of computer speech synthesis by lending it greater flexibility in speaking styles and expressions. In particular, we aim to use the methods developed and the knowledge gained in such investigations, ultimately to label the source units of a concatenative speech synthesiser with distinctive, emotion-related tags, thus allowing the synthesiser to choose more judiciously amongst the available units of the source database when required by the user to render a given text with a certain speaking style or emotion. This particular application will also provide a direct means of validating the auditory-perceptual significance of the proposed articulatory features or labels.

However, a number of limitations need to be overcome before deployment of the current methodology in speech synthesis. Perhaps the most fundamental limitation is the requirement of carefully measured formants, which are currently necessary in order to reliably estimate VT area-functions. A related issue is the reliance on formant bandwidths, after pre-treatment by statistical averaging and "closed-glottis" correction. In that regard, it is important to note that in the present study only a single set of bandwidths were used (i.e., those listed earlier in Table 1); it would be of particular interest to use one set of pre-treated bandwidths per emotion, and thereby to observe how the inclusion of emotion-related bandwidth variability may affect our present articulatory interpretations.

Another limitation of the current study is the use of only the steady-state segments of spoken vowels. In defence of that limitation, it is indisputably important to first gain an understanding of the basic structure and variability of the vowel space, whether it be the vowel space of particular languages, of particular speakers, or in the present case, of particular emotions. However, in line with much of the work on the acoustic-prosodic correlates of human vocal emotion, it would also be of significant practical and theoretical interest to extend our methodology to dynamic aspects of speech production, and thereby to model (and to segmentally label) emotion variability in articulatory gestures which span the space of the vowel targets studied here.

More generally, our segmental approach cannot be expected to render emotional speech single-handedly. Rather, a full realisation of speaking-style variation can be expected only in tandem with prosodic analyses. In that vein, we are in the process of considering our results in the light of F0 values measured in the same vocalic segments, and we also hope to use inverse filtering methods to investigate possible correlations of our results with laryngeal voice quality. It may also be revealing to investigate the potential dependence of our results on surrounding phonetic contexts, even though we sought to minimise such dependence by locating the steady-states of relatively uncoarticulated vowels.

Finally, the present study suffers a limitation which is common to most if not all studies in this area to date, i.e., the practical requirement to select a number of pre-defined emotions, in the face of a veritable explosion of humanly-possible expressions and speaking-styles. While the three emotional states Anger, Joy, and Sadness are amongst the so-called primary human emotions, in a recent review Cowie et al. (2001, p.36) state that "there is no agreement on a set of basic emotions", and even go so far as to state that "that lack of convergence suggests that there may well be no natural units to be discovered". It is our belief that ultimately, a more complete range of human emotions can be studied by collecting a sufficient range of natural, conversational speech data. The challenges and implications of such data collection are currently being pursued.



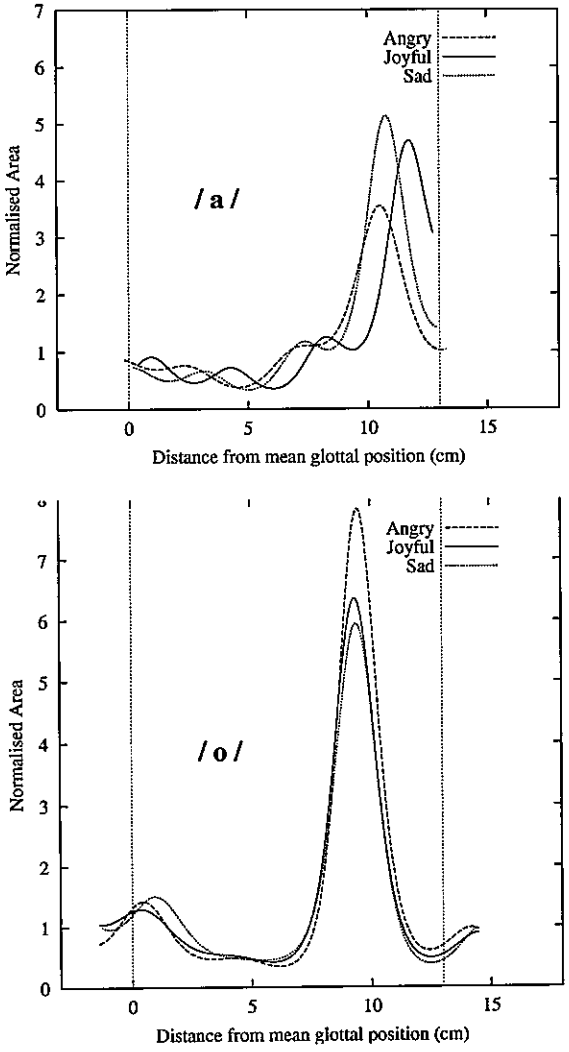


Figure 2. Emotion-related differences in vowel-specific articulatory *strategy*. The left and right, vertical lines in each panel mark the the mean position of the glottis (at 0cm) and of the lips (at 13cm), respectively.

REFERENCES

1. Burkhardt, F. & Sendlmeier, W.F., "Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis", in *Proc. ISCA Workshop on Speech and Emotion*, Newcastle, Northern Ireland: 151-156, 2000.
2. Campbell, N., "Limitations to concatenative speech synthesis", in *Proc. 6th Int. Conf. on Spoken Lang. Process.*, Vol. **III**: 416-419, 2000.
3. Campbell, W.N. & Black, A.W., "CHATR: Multi-lingual Speech Synthesis", *IEICE Technical Report* **SP96-7**, 1996.
4. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. & Taylor, J.G, "Emotion recognition in human-computer interaction", *IEEE Sig. Process. Magazine*, Vol.**18**, No.1, 32-80.
5. Erickson, D., Abramson, A., Maekawa, K. & Kaburagi, T., "Articulatory characteristics of emotional utterances in spoken English", in *Proc. 6th Int. Conf. on Spoken Lang. Process.*, Vol. **II**: 365-368, 2000.
6. Fónagy, I., "La mimique buccale: Aspect radiologique de la vive voix", *Phonetica* **33**: 31-44, 1976.
7. Iida, A., Campbell, N., Iga, S., Higuchi, F. & Yasumura, M., "Acoustic nature and perceptual testing of corpora of emotional speech", in *Proc. 5th Int. Conf. on Spoken Lang. Process.*: 1559-1592, 1998.
8. Kienast, M. & Sendlmeier, W.F., "Acoustical analysis of spectral and temporal changes in emotional speech", in *Proc. ISCA Workshop on Speech and Emotion*, Newcastle, Northern Ireland: 92-97, 2000.
9. Laver, J., *The phonetic description of voice quality*, Cambridge University Press, Cambridge, 1980.
10. Maekawa, K. & Kagomiya, T., "Influence of Paralinguistic Information on Segmental Articulation", in *Proc. 6th Int. Conf. on Spoken Lang. Process.*, Vol. **II**: 349-352, 2000.
11. Mokhtari, P., "An acoustic-phonetic and articulatory study of speech-speaker dichotomy", Doctoral Thesis, The University of New South Wales, Australia, 1998.
12. Mokhtari, P. & Clermont, F., "Parameters of unique vocal-tract shapes derived from linear-prediction of speech", (in preparation).
13. Mokhtari, P., Clermont, F. & Tanaka, K., "Toward an acoustic-articulatory model of inter-speaker variability", in *Proc. 6th Int. Conf. on Spoken Lang. Process.*, Vol. **II**: 158-161, 2000.
14. Murray, I.R. & Arnott, J.L., "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", *J. Acoust. Soc. Am.* **93**: 1097-1108, 1993.
15. Rank, E. & Pirker, H., "Generating emotional speech with a concatenative synthesizer", in *Proc. 5th Int. Conf. on Spoken Lang. Process.*: 671-674, 1998.
16. Snack software package, http://www.speech.kth.se/snack/.