J. Acoust. Soc. Jpn. (*E*) **21**, 4 (2000)

ACOUSTICAL LETTER

Static and dynamic vowels in a "cepstro-phonetic" sub-space

Frantz Clermont* and Shuichi Itahashi**

**School of Computer Science, University College, University of New South Wales, Canberra, ACT 2600, Australia*

***Institute of Information Sciences and Electronics, University of Tsukuba, 1-1-1, Tennodai, Tsukuba, Ibaraki, 305-8573 Japan*

(*Received 25 November 1999*)

Keywords: Vowels, Cepstrum, Formant, Isomorphism
PACS number: 43. 72. Ar, 43. 70. Hs

1. Introduction

Cepstral parameterisation of the speech signal has clearly become standard practice in the design of computer algorithms for spoken language recognition. Its successes stem mainly from the ability of the low-order cepstrum to reduce signal sensitivity to "noninformation-bearing variabilities",[1] thereby capturing the most distinctive features of speech spectra.[1,2] It has also become common practice to transform the low-order cepstrum on an auditory scale like the Mel, where low and high frequency ranges are respectively expanded and compressed.[3] This type of nonlinear transformation is expected to enhance phonetic distinctiveness manifest in gross spectral shapes,[1,2,4] which themselves are largely determined by the low-indexed, cepstrum coefficients.

An intriguing question then arises, whether certain of the low-indexed, Mel-cepstrum coefficients (*MCC*) effectively carry the *bulk of phonetic information*. In a succinct study of the Japanese monophthongs, Itahashi and Yokoyama[5] indeed identified the second (MCC_2) and third (MCC_3) coefficients as *likely correlates* of the two lowest formant-frequencies F_1 and F_2, respectively. However, the viability of the "cepstro-phonetic" sub-space thus implied, does depend on assembling evidence that is at least consistent for a range of vowel systems. In this paper, therefore, we report results obtained in a first step towards achieving that goal.

2. Speech corpus

In line with our quest for further evidence, the speech corpus selected for this study comprises English monophthongs and diphthongs, thus affording the possibility of observing static and dynamic behaviours of vowel cepstra for a language different from Japanese. The word list used for recording comprises 10 monophthongal syllables ("heed", "hid", "head", "had", "hard", "hud", "hod", "hoard", "hood" and "who'd") and 2 diphthongal syllables ("hay" and "high"). In a soundproof room, 5 adult-male and native speakers of Australian English uttered 5 random tokens of each of these syllables in citation style. The resulting, analogue signals were converted into digital form using a sampling frequency of 10 kHz and a quantisation precision of 12 bits.

3. Cepstral and formant parameterisation

All syllables thus digitised were subsequently analysed by linear prediction (LP-order 14) of Hamming-windowed frames (25.6-ms long), with a frame advance of 5 ms. Semi-automatic algorithms were employed to detect the 7 most stationary frames of the monophthongs' nuclei,[6] and all voiced frames of the diphthongs' nuclei.[7] At those frames, 14 LP-cepstrum coefficients were extracted initially on linear Hertz-scale, and then warped to match the Mel-scale using the Bilinear Transform[8] and a warping factor adjusted to 0.34 for the sampling frequency of 10 kHz in effect. Across the same frames, the 4 lowest formant-frequencies (F_1, F_2, F_3 and F_4) were estimated by temporally constrained, analysis-by-synthesis.[7]

4. Cepstro-phonetic sub-space

In Section 4.1, a phonetically motivated search for the most influential *MCC*-pair yields results that provide quantitative support for Itahashi and Yokoyama's[5] earlier findings. In Section 4.2, the notion of a cepstral sub-space isomorphic to the F_2-F_1 plane is reinforced through a number of compelling observations on the distributions of monophthongs and diphthongs in that sub-space.

4.1 Phonetically motivated search

Only the monophthongs were implicated in the search, and first divided into 4 broad classes described in articulatory phonetics as {high} versus {low} and {back} versus {front}. This dual, binary classification was carried out by simple thresholding about the 2 lowest formants of a neutral vocal-tract: {high} if $F_1 <$ 500 Hz and {low} otherwise; {back} if $F_2 < 1{,}500$ Hz and {front} otherwise. The {back}-{front} sets obtained comprise *MCC*'s for {"hard", "hud", "hod", "hoard" and "hood"}-{"heed", "hid", "head", "had" and "who'd"}. The {high}-{low} sets retained include *MCC*'s for {"heed", "hid", "hoard", "hood" and "who'd"}-{"had", "hard", "hud" and "hod"}. These results were achieved unanimously for all speakers, and for all vowel instances (implying the 7 steady-state

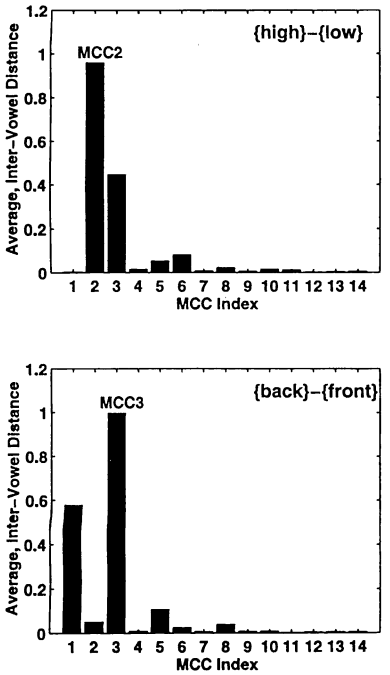

Fig. 1 Profiles of Euclidean distances indicating the average dispersion, per Mel-Cepstrum Coefficient (MCC), for {high}-{low} and {back}-{front} classes of Australian English monophthongs, and yielding MCC_2 and MCC_3 as the most influential MCC-pair (see Section 4.1 for further details).

frames and the 5 tokens) pertaining to the monophthongs classified above.

The search procedure itself consisted of computing, for each MCC and each speaker, the Euclidean distances between vowel instances in the {high}-set and vowel instances in the {low}-set, and then averaging over all such distances and all speakers. By repeating this procedure for the vowel instances in the {back} and the {front} set, two profiles of distances shown in Fig. 1 were obtained against MCC-index on the horizontal axis. The top graph indicates that maximum {high}-{low} dispersion is achieved with MCC_2, while the bottom graph indicates that maximum {back}-{front} dispersion is achieved with MCC_3. Both graphs also show second-candidate coefficients, whose contributions appear to be non-negligible and therefore are worth further investigation. This notwithstanding, Fig. 1 provides some empirical justification for claiming that, on average, MCC_3 and MCC_2 carry the bulk of distinctive information in vowel spectral shapes. As a first step towards verifying this claim, however, it will be reassuring to determine at least visibly that, by analogy with the F_2-F_1 plane, the MCC_3-MCC_2 plane is able to yield vowel (static and dynamic) distributions

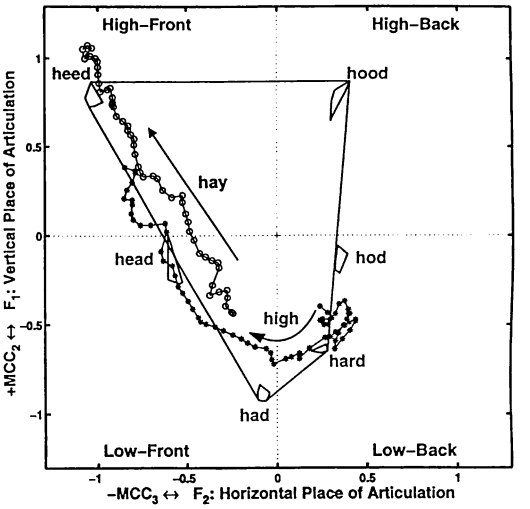

Fig. 2 A cepstral homologue of the F_2-F_1 plane: $-MCC_3$ on the horizontal axis and $+MCC_2$ on the vertical axis (see Sections 4.1 and 4.2 for further details).

that are consistent with basic notions of articulatory phonetics.

4.2 Some validating observations

The cepstral plane (MCC_3-MCC_2) targeted above is shown in Fig. 2 for the 2 diphthongs and a subset of the monophthongs considered in this study. A convincing portrayal of the time-honoured quadrilateral is first observed around the so-called cardinal vowels in "heed", "had", "hard" and "hood", which were produced by one of our 5 speakers. It is interesting that the vertices of the quadrilateral should be conveniently located in distinct quadrants of the plane, where MCC_3 and MCC_2 coordinates carry algebraic signs that appear to cue broad phonetic distinctiveness at least for the speaker's data illustrated in Fig. 2. Consistent with this sign polarity are the vowels in "head" and "hod" produced by the same speaker, which are located at about midway between high and low vowels and whose MCC_2 values are clustered in the neighbourhood of the zero line.

The MCC_3-MCC_2 plane also appears to render the dynamics of certain diphthongs without distortions. Prototype trajectories through the nuclei of "hay" and "high", uttered by the same speaker, are superimposed in Fig. 2 and, indeed, are seen to follow trans-monophthongal paths whose respective direction and overall non-linearity agree well with those observed in formant space.[7,9,10] Furthermore, the cepstro-temporal movements through the diphthongs unfold either near the boundaries of or within the speaker's monophthongal quadrilateral. This type of consistency does lend some strong credibility to the claimed isomorphism between the F_2-F_1 and the MCC_3-MCC_2 plane.

5. Conclusion

Using static and dynamic vowels in spoken Australian English, we have presented some quantitative evidence reinforcing the existence of 2 cepstral dimensions that embody basic properties attributed to the 2 lowest formants. Of the 14 lowest, Mel-cepstrum coefficients investigated in a phonetically motivated search, the third (MCC_3) and the second (MCC_2) were found to yield, on average, the greatest {back}-{front} and {high}-{low} separation of steady-state monophthongs, respectively. Further evaluation of the MCC_3-MCC_2 plane was also attempted by examining therein the diphthongs' trajectories in "hay" and "high" and confirming their expected movements through the steady-state monophthongs. Both static and dynamic vowels have MCC_3 and MCC_2 coordinates whose algebraic signs appear to be distinctive and thus could be said to carry some basic articulatory-phonetic cues.

In sum, the empirical results reported above for spoken Australian English confirm and extend Itahashi and Yokoyama's[5] seminal observations recorded for spoken Japanese. In this sense, therefore, the evidence accrued to date tends to support the notion of "cepstro-phonetic" sub-space advanced here to describe the isomorphism observed between the F_2-F_1 and the MCC_3-MCC_2 plane. However, a number of questions are still unanswered, which will require further but interesting research. The situation depicted in Fig. 1, for example, invites further exploitation of the theoretical formula[3] relating LP-cepstrum to formants with a view towards exposing the intrinsic behaviour of MCC_3 and MCC_2 in the frequency ranges of F_1 and F_2. The viability of the MCC_3-MCC_2 plane will also depend upon achieving consistency and robustness, which need to be addressed across a wider range of sound systems, and a variety of speakers and recording conditions. As these questions are progressively elucidated, the prospect of utilising a cepstro-phonetic space for cross-language comparison by computer will become more tangible.

Acknowledgments

The research work described in this paper was conducted in 1998-1999 during the first-author's sabbatical visit to the University of Tsukuba (College of Information Sciences). Grateful thanks are extended to Professor Naoto Sakamoto for his encouragement and to Drs Parham Mokhtari, Michael Barlow and David J. Broad for their useful comments.

References

1) L. Rabiner and B-H. J. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, 1993), p. 169.
2) S. Furui, "Research on individuality features in speech waves and automatic speaker recognition techniques," Speech Commun. **5**, 183-197 (1986).
3) S. Itahashi, "On properties of speech cepstra," English version of a paper published in Trans. Inst. Electron. Inf. Commun. Eng. **J71-D**, 1839-1842 (1988) (in Japanese).
4) H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am. **87**, 1738-1752 (1990).
5) S. Itahashi and S. Yokoyama, "Feature extraction of speech by modified cepstrum analysis," English version of a paper published in Proc. Spring Meet. Acoust. Soc. Jpn., 237-238 (1974) (in Japanese).
6) P. Mokhtari, "An acoustic-phonetic and articulatory study of speech-speaker dichotomy," Unpublished PhD Thesis, University of New South Wales (1998).
7) F. Clermont, "Formant-contour models of diphthongs : A study in acoustic phonetics and computer modelling of speech," Unpublished PhD Thesis, Australian National University (1991).
8) J. W. Picone, "Signal modelling techniques in speech recognition," IEEE Proc. **81**, 1215-1247 (1991).
9) H. Piir, "Acoustics of the Estonian diphthongs," in *Estonian Papers in Phonetics* (Academy of Sciences of the Estonian SSR, Tallinn, 1982-1983), pp. 5-103.
10) F. Clermont, "Spectro-temporal description of diphthongs in F_1-F_2-F_3 space," Speech Commun. **13**, 377-390 (1993).