RECOGNITION OF CHINESE WORD TONE FROM FO, WITH AND WITHOUT AMPLITUDE AND
SPEAKER INFORMATION

J. H. Leather *

INTRODUCTION

Many of the problems of intonation recognition are encountered also (and perhaps in more tractable form) in the domain of lexical tone - the syllable-level prosodic patterning which in many languages distinguishes between words with identical segmental structures. For 'tone' languages - which may even be in a majority (ref 7, 9) - tone recognition is necessary for lexical disambiguation. Moreover, when tone interacts with intonation (ref 5), factoring out one may be a prerequisite for the extraction of the other. In this paper, procedures are described for the recognition of tone in citation-form monosyllables of Mandarin, i.e. Modern Standard Chinese (MSC).

THE MSC TONE SYSTEM

A syllable in MSC includes in its phonological specification one of four pitch patterns or 'tones' which establish lexical contrasts between words with otherwise identical phonological forms (ref 3). Tone thus functions at the syllable level to reduce by an order of magnitude the number of homophones in the dictionary. For example, the monosyllable /y/ with Tone 1 is a word meaning 'muddy'; with Tone 2 it is 'fish'; with Tone 3 'rain'; and with Tone 4 'jade'.

Auditorily, the tones in citation form are analysable as patterns of voice pitch, and are often represented as trajectories at or between pitch levels on a five-point scale. The tones in this notation are: 4-4 (Tone 1, level); 2-4 (Tone 2, rising); 2-1-3 (Tone 3, falling-rising) and 5-1 (Tone 4, falling). Instrumental studies (ref 2, 6, 8) have shown the tones to be realized with characteristic FO contours which broadly correspond to the traditional auditory pitch analyses. Figure 1 shows FO contours of four tonally-distinguished monosyllabic words produced by a male native speaker of MSC.

PATTERN ELEMENTS AND SPEAKER ADAPTATION IN TONE RECOGNITION

FO patterning is the primary perceptual cue for tone in MSC (ref 12, 15), but accurate automatic tone recognition may not always be possible from FO contour shape alone. Cheng and Sherwood (ref 4) report correct recognition from FO contour shapes of 90% of tonal monosyllables spoken by the male speaker to whose speech the parameters of their recognition routine were set, and 84% of the same items produced by another male speaker. However, their procedure may be less successful with the perceptually equivalent FO contours of tone productions by speakers with very different voice ranges - particularly female or child speakers - as it bases decisions upon absolute rather than relational (multiplicative) differences between values of FO at sampling points in the test utterances.

Perceptual studies (ref 12) have indicated that apart from the shapes of

* University of Amsterdam, Spuistraat 210, 1012 VT Amsterdam, Netherlands
the F0 contours, the tones are cued also by the F0 levels of the contours, and by patterns of intensity (see also refs 13, 14). F0 levels are interpreted in perception by reference to some actual or supposed speaker voice range (ref 10), but how such F0 calibration should best be achieved in automatic tone recognition remains an open question. In the recognition experiments described below, possible contributions of both amplitude and speaker information to recognition accuracy are examined.

EXPERIMENTAL CORPUS

The tone productions based on which the recognition algorithm was constructed were provided by three male and three female speakers in their twenties who had been raised in Beijing (though in one case originating in Shanghai) and were on a study visit to England. Each spoke (from cards, in random order) a total of three tokens of each of the words /ti/, /di/ and /y/ in each of the four tones, and the words /k~I/, /g~I/ and /da/ in Tone 1. Besides the acoustic pressure waveform, the Lx output of a laryngograph was recorded for subsequent analysis of time-varying F0. Of the total of 270 tonal monosyllables, 8 subsequently proved to be unusually noisy. To provide data for the computation of individual voice statistics, each informant also read aloud an excerpt from a dramatic narrative.

RECOGNITION FROM FO CONTOUR SHAPE

A tone recognition algorithm was designed to look at three points in the F0 contour of an utterance, and base an identification decision on their interrelationship. These points were at 20%, 50% and 85% of the duration of voicing in the whole utterance. (Sampling at 20% and 85% duration, rather than at the beginning and end, effectively launders the F0 contour of the fine-detail variations and irregularities due to post-obstruent onset, offset and the characteristic initial 'ramp' of a citation form.)

The algorithm assesses the slopes of the bi-linear stylized curves to which the F0 contour is thus reduced, making a tone identification accordingly. The algorithm was implemented on a microcomputer interfaced to a Voiscope, using routines for larynx period sampling and vetting which have been described elsewhere (ref 11).

This basic algorithm correctly identified 247 out of the 262 usable F0 contours. Analysis of the recognition errors showed that all but two involved mistaking Tone 3 for Tone 2 or vice versa, and a re-examination of the misidentified F0 contours indicated that these confusions were due to inter-speaker variation. It was found that correct recognition of two of the misidentified tokens could be achieved by taking the mid-contour F0 value not always at 50% of the duration of voicing, but at that point between 25% and 65% duration where F0 was lowest. It appeared from examination of the other misidentified tokens that the remaining Tone 2 / Tone 3 confusions might be resolved by introducing a speaker-adaptation factor.

SPEAKER-ADAPTATION I: LONG-TERM SPEAKER VOICE STATISTICS

An F0 reference level was determined for each of the informants as follows. The 8-bit larynx period values of each speaker were sampled at 10 msec intervals during the text reading to provide a total of around 2500 observations. The frequency distribution of these observations was third-order mean smoothed, and the logarithmic mean of the values
corresponding to the upper and lower 1% limits of the distribution was taken as a voice mid-range reference value. The Tone 2 and Tone 3 tokens in the corpus were then examined to determine, speaker by speaker, the relationship of the FO at 50% duration to this reference value. It was found that the relationship was consistent enough to distinguish between the two tones. Incorporating this speaker-adaptation factor and some corresponding minor adjustments, the algorithm was capable of correctly recognising all the tone tokens except one.

SPEAKER-ADAPTATION II: TONE 1 AS REFERENCE LEVEL

Although the FO range spanned by the tones varies with affect (ref 1), systematicity is maintained by a constant interrelationship between the FO contours of the individual tones, so that one 'reference' tone may be theoretically capable of providing the calibration information necessary for correct recognition of the others in the system. This was an alternative approach to speaker-adaptation. Mid-contour FO values were determined for each speaker's tokens of Tone 1 (the level tone), and mid-contour values of Tone 2 and Tone 3 tokens compared with this Tone 1 reference level. It was found that with one exception (the same production by the Shanghai-born speaker as was misidentified using long-term speaker voice statistics) Tone 2 and Tone 3 were distinguished, quite apart from their differences in FO contour shape, by their respective percentage deviations from the mean Tone 1 value. With this means of speaker-adaptation incorporated in the algorithm, recognition of the corpus items could again therefore achieve more than 99% accuracy.

SPEAKER-INDEPENDENT RECOGNITION FROM COMBINED FO AND AMPLITUDE

The last experiment investigated the use of amplitude as a secondary parameter which might permit more accurate speaker-independent recognition. As in previous studies (e.g. ref 8) it was found that smoothed RMS amplitude contours in most cases differed distinctly between Tone 2 and Tone 3. Only the latter had a dipping contour. Adding the criterion of whether amplitude at 15% duration was greater than at 40% (= Tone 3) enabled the algorithm to decide correctly between all but three of the problem cases.

CONCLUSION

These procedures evidently need to be tested on a larger body of data from new speakers (and with a greater proportion of Tone 2 and Tone 3 tokens). The present results, however, suggest that MSC citation-form tones can to a useful extent be recognised on the basis of FO contour shape alone, with only Tones 2 and 3 presenting some problems. For reliable recognition of such monosyllables in the absence of prior speaker information, amplitude can be used to help distinguish between the problematic Tones 2 and 3 until a stable enough mean FO value for an individual speaker’s Tone 1 is available to provide the necessary FO reference level for speaker-adaptation.
Figure 1

Smoothed time-normalized F0 of /ti/ in Tones 1, 2, 3 and 4 (one speaker)

REFERENCES