Voice characteristics of spontaneous speech

Nick Campbell
ATR Human Information Science Laboratories,
Keihanna Science City, Kyoto, Japan
nick@atr.co.jp

1 Introduction

There has been considerable interest recently in the characteristics of spontaneous speech. This paper examines some acoustic characteristics from a very large corpus of interactive conversational speech and reports findings which show that voice quality, and corresponding spectral tilt, is varied consistently, in much the same way as pitch and duration, and that it correlates well with perceptual classifications of paralinguistic features in the speech.

2 What is spontaneous speech?

In the context of speech technology research, the term 'spontaneous-speech' has traditionally been used in contrast to 'read-speech' as an indicator of the degree of control in speech utterance production, according to whether the speech content is generated in real-time, while speaking, or is simply converted from text through a process of reading. Spontaneous speech is thought to be 'more noisy' (hesitations and fillers) and 'less-well-formed' (or un-grammatical). However, it is an over-simplification to assume that spontaneity is a binary attribute of speech. There are degrees of spontaneity, even in read speech, and speaking styles can vary in a range between the highly-rehearsed formal presentation style (e.g., for broadcasting and public-speaking), and the intimate chatting betweeen friends and family members. With the former, the controlled structure of the speech arises from a predominance of lexical information (and often by a reliance on a written text as the original basis for the speech), but with the latter, the degree of shared common-knowledge is much higher, and much of the spoken interaction takes place in a non-verbal form. Often its purpose is not to impart information, but simply to be social.

The JST/CREST ESP Corpus [1] exemplifies the latter. It consists of wholly unprepared speech, with controls for the degree of familiarity between speaker and hearer. In this paper, we present results of an analysis of part of this corpus, showing that the same lexical string, spoken by the same speaker, often carries different paralinguistic information, and we confirm that independent listeners can form a similar context-independent interpretation of this 'meaning-behind-the-words' from similarities in the prosodic and voice-quality parameters.

The biggest difference that we notice between this corpus and others that are currently available lies in the amount of phatic communication, or 'interactive social speech'. People speak not to negotiate information, but rather to express relationships. This often takes the form of 'back-channeling' and 'fillers' (sounds which are currently regarded as 'noise' from the point-of-view of speech-procesing techniques). We believe that this aspect of speech communication might be useful for both recognition and synthesis technologies to enable some 'reading-between-the-lines' in processing conversational or interactive speech information.

3 Acoustic features

The software used in this work is available from the ESP web-site [2]. It consists of a graphical interface (written in tcl/tk) for displaying subsets of the corpus as points in a space for labellers to re-arrange into groups having similar perceptual aspects. Figure 1 shows a sample screen; the meters on both sides of the screen are not displayed for the labellers, but present the acoustic characteristics of each data point in simple visual form for subsequent checking of the resulting groupings by speech researchers.

The feature extraction uses simple routines provided by the Snack sound library distributed by KTH [3]. The meters on the left (Fig.1) show F0 mean, max, and min in relative terms, defined by the observed distributions of the subset of the data currently being displayed. The bottom 3 meters show degree of voicing, position of the F0 peak, and position of the rms-amplitude peak for the utterance pointed to by the cursor.. The meters on the top-right show mean, max, and min for rms amplitude, and those below show duration and two measures of spectral tilt (H1-H2, and H1-A3, as used by Hanson [4], Sluijter [5], and the present author in previous work [6]). We have proposed an improved method for measuring voice-quality [7], but it is not yet incorporated in the present software.

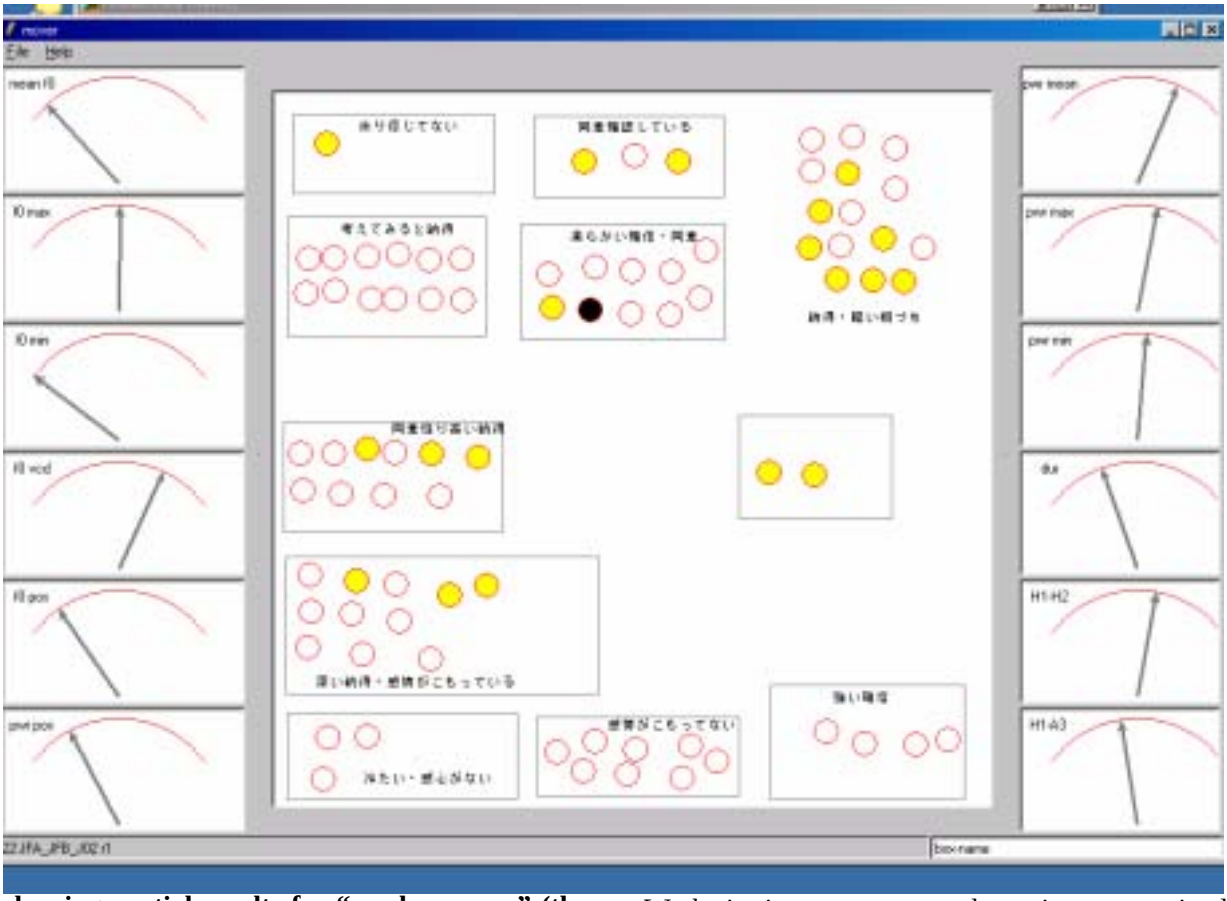
When labellers first open the softeware, they select a word or phrase that appears more than about 75 times in the corpus from a given speaker, and the points representing each utterance of that phrase are aligned along the main diagonal in order of their appearance in the corpus. By clicking on each point, the labellers can hear the phrase (as many times as they like) and are able to move it to a different place on the screen. They are free to form as many groups as they wish, and can surround each group with a box having a label which they are free to determine. In this way, similar to sorting a record collection, the categories emerge from the data.

Subsequent post-processing (illustrated by the meters) determines the acoustic categories of each group. In order to determine the priorities among the current 12 features, a classification tree [8] is grown and pruned back for robust prediction. The surviving nodes on the tree indicate the strongest features correlating with each group. Further work is being carried out to determine the influence of dynamic speech characteristics, but the present paper presents results only from the simpler static features, as an illustration.

Figure 1. The perceptual labelling software,

自然対話発話の声質特徴

○ ニック キャンベル
ATR-人間情報科学研究所、けいはんな学研都市、京都 .



showing partial results for "soudesuyoune" (the meters are not usually shown while labelling)

4 Paralinguistic features

The labels that emerged for the phrase "soudesune" included: - 意外性の強い納得 / 感情がこもってない / 強い確信 / 強い確信自信がある / 考えてみると納得 / 柔らかい確信・同意 / 深い納得・感情がこもっている / 同意確認している / 同意性より高い納得 / 納得・軽い相づち / 余り信じてない / 冷たい・関心がない , etc., reflecting the different pragmatic force of each utterance type.

The four different labellers did not use the same categories or the same number of groupings, making direct comparison of the results difficult, but a tree grown and pruned for each data set using the public-domain 'R' statistical package [9] showed the following four acoustic features to be consistently distinctive for each labeller from among the twelve features currently being measured: f0.mean/min: mean and lowest measured pitch in the utterance, f0.pos: relative position of the F0 peak in the utterance, and h1.a3: spectral tilt of the utterance.

Table 1. R-Tree output for one labeller's dataset:

>summary(tree(formula = factor(percept) ~ h1.a3 + h1.h2 + f.mean + p.mean + f.pos + f.vce + f.min + p.min + dur))

Variables actually used in tree construction:

[1] "f.min" "f.pos" "h1.a3" "f.mean"

Number of terminal nodes: 10

Misclassification error rate: 0.2968 = 19 / 64

5 Discussion

Work is in progress to determine an optimal granularity for labelling affect in speech. Listeners appear consistent in their categorisations, but there are many different aspects of non-verbal information and we do not yet have a clear framework for their description. For this paper, we suffice to note that the categorisations within labeller can be well modelled by a small number of acoustic parameters, and note especially that in all cases so far examined, the spectral-tilt (i.e., the tense-breathy dimension) appears to be a strong predictor. This confirms our earlier findings [10] that voice-quality functions as a prosodic parameter.

Acknowledgements

The author is grateful to the JST-CREST and members of the ESP project for help with the production and analysis of this data.. This work was also partly supported by the Telecommunications Advancement Organisation of Japan.

References

[1] [2] ESP pages: http://www..feast.his.atr..co.jp
[3] Snack : http://www.speech.kth.se/snack
[4] Hanson H., Unpublished PhD thesis, MIT, 1997.
[5] Sluijter, A.,
[6] Campbell, N.,& Beckman, M.,
[7] Mokhtari, P. & Campbell, N. (**2002**). "Automatic characterisation of quasi-syllabic units for speech synthesis based on acoustic parameter trajectories: a proposal and first results", in *Proc. Autumn-02 Meet. of the Acoust. Soc. Japan*, Akita, 233-234
[8] Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984) Classification and Regression Trees. Wadsworth..
[9] Ihaka, R., and Gentleman, R., "R: A Language for Data Analysis and Graphics", Journal of Computational and Graphical Statistics, vol5.3,pp.299-314, 1996
[10] Campbell & Mokhtari, Voice Quality, the 4th prosodic dimension, in Proc ICPhS 2003, Barcelona.