Listening between the lines: a study of paralinguistic information carried by tone-of-voice

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

1. Introduction

This paper describes a study of speaking-style characteristics, or "tone-of-voice", in conversational speech, and shows that non-verbal information is transmitted efficiently regardless of cultural and linguistic contexts through differences in prosodic and voice-quality features. We asked Korean and American listeners with no previous kowledge of Japanese to judge the meaning of various utterances of the interjection "eh" in Japanese conversations, and compared their perceptions with the acoustic characteristics of the speech. We found voice- quality information to be an important discriminator.

2. Paralinguistic speech characteristics

Speech research has moved from a phase in which it was concerned predominently with the study of lab-speech, or read speech, into one in which it focuses more on spontaneous speech. The latter is often thought to be 'noisy' or 'ill-formed', as a result of its many hesitations and fillers. However, we claim that these 'noises' (or non-verbal speech sounds) are an important part of inter-personal communication and that they serve to display affect and discoursal information as much as lexical and syntactic content conveys propositional information. The controlled structure of lab-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 conversational speech, the degree of shared knowledge between speaker and listener is much higher, and as a consequence, a large part of the spoken interaction takes place in a non-verbal form. Often the purpose of such speech is not to impart information, but simply to be social.

The JST/CREST ESP Corpus [1,2] consists of wholly unprepared speech, with labels for the degree of familiarity between speaker and hearer, and for discoursal and affective functions. In this paper, we present results of an analysis of part of this corpus, showing that the same lexical string, a word spoken by the same speaker, often carries different paralinguistic information. We show from the results of our analysis 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 only negotiate information, but rather to express relationships [3]. This often takes the form of 'back-channeling' and 'fillers' (sounds which are currently regarded as 'noise' from the point-of-view of speech-processing, and therefore disregarded, but we believe that this aspect of speech communication might be useful for both recognition and synthesis technologies to enable 'reading-between-the-lines' when processing conversational or interactive speech signals.

3. Data for the analysis

We selected 129 utterances of the word "eh" from this large Japanese spontaneous-speech corpus, and asked listeners in Korea and America to judge the meaning or intended effect of each utterance using software specifically designed for sorting and categorising speech tokens ('Mover' [4], see below). The listeners were asked to indicate for each "eh" utterance which of the following set of descriptors it best approximated: *[annoyed, angry, confused, excited, happy, hello, hurt, indifferent, indifferent/uncaring, relieved, sad- crying, sad, scared, sick-of, surprised, uncaring, unsure]*

The list was generated from a larger list of about 85 adjectives suggested by Japanese listeners using an open-input version of the same 'mover' software. The utterances were listened to in isolation, with no previous or following discourse-context information provided. Listeners were free to listen to each utterance as many times as they felt necessary and were asked to group the utterances into the above categories as they felt appropriate [5].

This previous work compared the responses of the non-Japanese listeners (who had no previous familiarity with the Japanese language) with those of the native Japanese respondents, and showed that although there is considerable individual variation in the choice of descriptor, all listeners were able to perceive broadly similar effects, and a principal component analysis identified the dominant dimensions as fitting well into the valency- activation framework described in the psychological literature [6]._The present paper looks at the acoustic characteristics of these speech utterances, in the context of the valency-activation dimensions and attempts to explain the listener's responses.

4. Acoustic features

The software used for this work is available from the ESP web-site [7]. 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 utterance in simple visual form for subsequent manual checking of the results by speech researchers and for output to a data file for the utterance set.

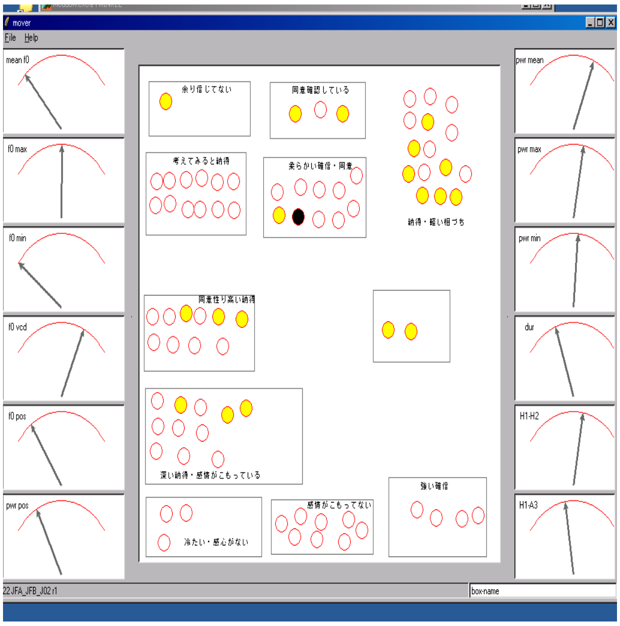

Figure 1. *The perceptual labelling software, showing partial results for utterance categorisation (the meters are not usually shown while labelling)*

The feature extraction uses simple routines provided by the Snack sound library distributed by KTH [8]. The meters on the left show F0 mean, maximum, and minimum in relative terms, i.e., normalised between zero and one, as defined by the observed distributions of the subset of the data currently being displayed. The bottom three meters show degree of voicing, relative position of the F0 peak, and relative position of the rms-amplitude peak for the utterance pointed at by the cursor. The meters on the top-right of the figure show mean, maximum, and minimum for rms amplitude, and those below show duration and two measures of spectral tilt (H1-H2, and H1-A3, as used by Hanson [9], Sluijter [10], and the present author in previous work [11]). We have proposed an improved method for measuring voice-quality [12], but it is not yet incorporated in the present software.

When labellers first open the software, the speech samples are represented as small circles 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 then surround each group with a box having a label to be selected from the above list. Boxes can overlap or be nested.

The output from the labelling program was stored as text files, indicating the category determined for each speech token. We performed a Principal Component Analysis ('princomp' in "R" [13]) on both the perceptual category data and on the acoustic characteristics of the speech utterances. Results of the perceptual analysis were presented in [5], and we focus here on linking these to the acoustic data analysis. Figures 2 and 3 show bi-plots of the perceptual labels (listed with translations in Table 1) with arrows indicating the strength of each feature in this view of the space. The two components account for 32% and 25% of the variance respectively.

Table 1. *The labels for the PCA plots of perception:*

ead	annoyed	kgc	guichanta (annoyed)
ean	angry	khn	hwaganada (angry)
eco	confused	kch	chilmunhada (question)
eex	excited	khd	haengbunhada (excited)
eha	happy	khb	haengbokhada (happy)
ehe	hello	kay	anneyong (hello)
eht	hurt	kta	tachida (hurt)
ein	indifferent	kda	darum (other)
eiu	indiff/unc	km2	musimhan (indifferent/uncaring)
erl	relieved	kas	ansimhada (relieved)
esc	sad-crying	ksp	sulpuda (sad/crying)
esd	scared	km3	musupda (scared)
eso	sick-of	kap	apudan (sick-of)
esp	surprised	knl	nolada (surprised)
eun	uncaring	kmu	musimhada (unconcerned)
eus	unsure	kpl	pulanjohghan (unsure)

5. Principal component findings

From the principal component analysis of the perceptual results, we found that both Korean and American listeners were able to distinguish differences between the 129 utterances of "eh" and were able to align them similarly in terms of (for the first two components at least) valency and activation, the two main dimensions that have been suggested in the psychological literature as explaining expression of emotion or affect.

We can see from figures 2 and 3 that the descriptors are aligned similarly, although the plots show the space to be rotated. It is known that the directions of the axes in a principal component analysis are arbitrary, being unitless but scaled equivalently, so the sign of these axes can be ignored, showing the two plots to be equivalent. The vertical axis in both plots probably represents activation. It is not always clear in a principal component analysis what each dimension might represent, but strong descriptors are found at the top of each plot and weak ones at the bottom. The horizontal axis probably represents valency, with positive descriptors on the right for the American responses and on the left for the Korean ones, and negative descriptors on the left for the American and on the right for the Korean responses (for a more detailed discussion of these results and their implications, please refer to [5]).

Table 2. *Loadings from the principal component analysis of acoustic features. The bottom line shows the cumulative percentage of variance explained by each component. Bold figures indicate the dominant features in each dimension.*

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
fmean	**-0.381**	-0.288	-0.223	-0.170	0.146
fmax	-0.369	-0.189	-0.230	-0.221	---
fmin	-0.235	**-0.471**	-0.190	-0.123	0.269
fvcd	-0.334	0.230	0.132	-0.158	-0.200
fpos	-0.244	---	---	-0.312	**-0.567**
ppos	0.124	-0.232	0.213	**-0.503**	-0.435
pmean	-0.330	0.341	0.235	---	0.123
pmax	-0.376	0.152	---	---	0.224
pmin	---	---	**0.649**	-0.172	**0.338**
dur	-0.116	0.347	**-0.474**	0.142	-0.134
h1h2	-0.285	-0.174	0.264	**0.468**	-0.294
h1a3	---	**-0.488**	0.104	0.416	-0.215
a3	**0.345**	-0.150	-0.118	-0.288	0.124
percent:	40.6	54.1	65.2	74.7	82.8

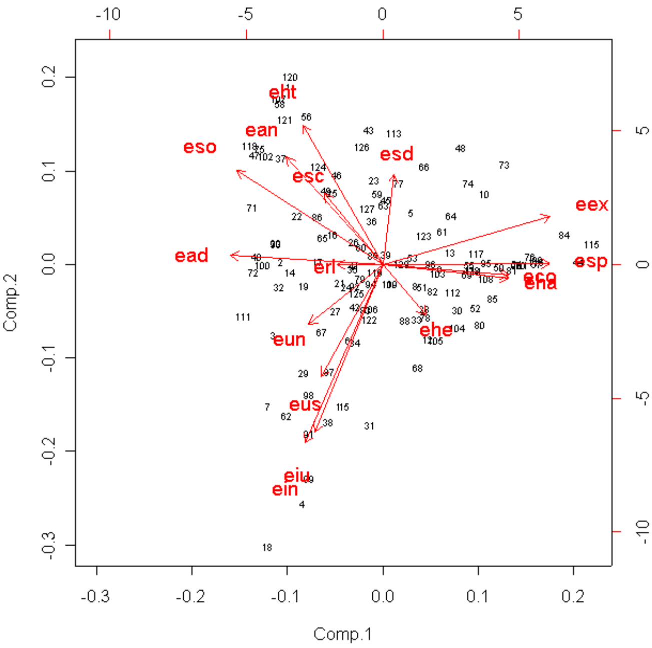

Figure 2. *The first two dimensions of the principal component analysis for the American listeners' perceptions*

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
ead	**-0.364**	---	0.211	-0.301	---
ean	-0.230	0.293	**0.343**	0.227	0.271
eco	0.301	---	0.334	---	-0.231
eex	**0.400**	0.132	---	---	0.148
eha	0.302	---	---	**-0.418**	---
ehe	0.104	-0.138	-0.101	-0.233	**0.548**
eht	-0.190	**0.381**	---	0.158	0.250
ein	-0.185	**-0.487**	---	---	0.161
eiu	-0.162	-0.457	---	---	0.172
erl	-0.109	---	-0.183	-0.270	0.216
esc	-0.140	0.193	-0.418	-0.323	**-0.418**
esd		0.247	**-0.579**	0.228	---
eso	-0.349	0.256	0.287	-0.133	---
esp	0.399	---	0.230	---	---
eun	-0.176	-0.164	---	-0.206	-0.408
eus	-0.147	-0.305	-0.114	**0.541**	-0.176
percent:	17.8	32.1	41.5	49.5	57.3

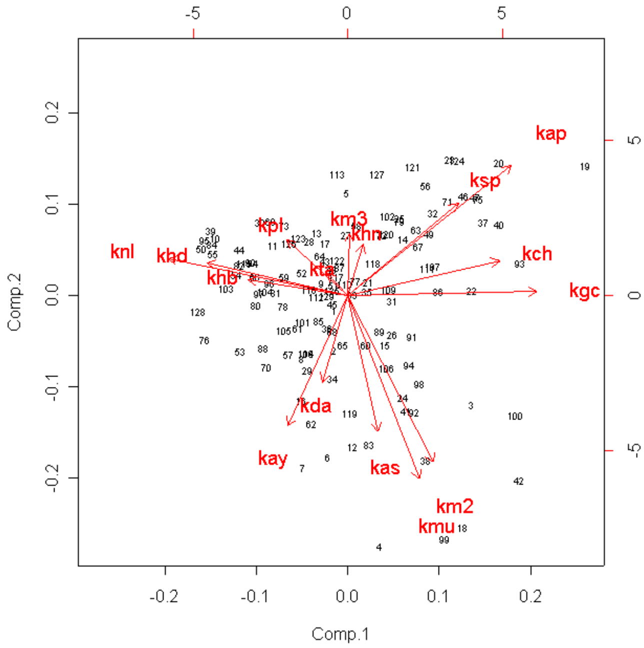

Figure 3. *The first two dimensions of the principal component analysis for the Korean listeners' perceptions*

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
kap	0.383	**0.342**	0.129	---	---
kas	---	-0.359	0.206	0.254	**0.357**
kay	-0.140	-0.344	0.185	0.371	0.341
kch	0.357	---	**0.278**	0.131	---
kda	---	-0.228	**-0.433**	**-0.382**	0.191
kgc	**0.445**	---	---	0.126	---
khb	-0.233	---	0.371	-0.124	0.312
khd	-0.330	---	---	---	-0.153
khn	---	0.135	-0.421	**0.536**	0.181
km2	0.201	-0.440	---	-0.144	-0.342
km3	---	0.165	-0.257	0.329	-0.307
kmu	0.168	**-0.483**	---	---	-0.275
knl	**-0.420**	---	0.156	---	**-0.362**
kpl	-0.144	0.147	0.207	-0.171	---
ksp	0.259	0.243	0.139	-0.217	0.172
kta	---	---	-0.401	-0.295	0.313
percent:	14.3	25.7	34.6	42.3	49.8

We performed a similar analysis of the acoustic features of each utterance using the feature descriptors produced by the 'Mover' software and the normalised values for each acoustic component listed above. Figure 4 plots the first two components and figure 5 plots the third and fourth. We can see from Table 2, which shows the loadings of each factor in this analysis, that the first four components account for approximately 75% of the overall variance, and the first two explain more than 50%.

From our previous work, we had expected to find that the first dimension fits fundamental frequency, the second dimension differences in duration, the third power, and the fourth spectral-tilt. However, inspection of Table 2 reveals a different ordering. We can see that the first dimension does indeed fit well to differences in fundamental frequency, but the next three reveal interesting information. It appears that voice-quality is more important.

The *a3* feature (amplitude of the third formant) in component 1 accounts well for a small cluster of creaky-voice utterances that have no energy in the area of the third formant because they are unvoiced. This appears to be a stronger indicator than *fvcd* (degree of voicing). However, the second component features *fmin* and *h1a3* most strongly. These reflect voice-quality. Our previous work has shown this dimension to vary consistently in discourse-related ways [3], but the present analysis might be interpreted as indicating that voce quality has a more important role in portraying paralinguistic information in non-lexical utterances. The third dimension includes duration and minimum power, both indicating strength of utterance, and the fourth includes both *ppos* and *h1h2*. Both these features represent amplitude, or power of the voice. *h1h2* is a simple measure of spectral energy distribution, and ppos indicates the position in time of the peak of maximum energy in the utterance.

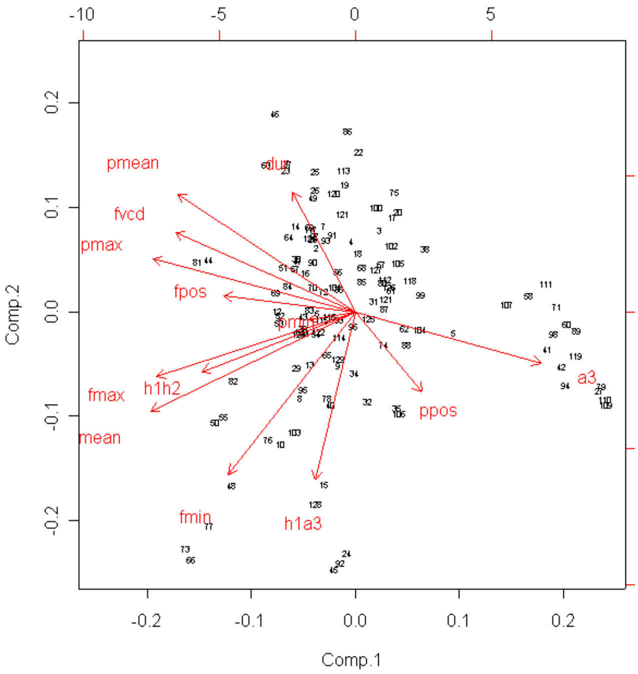

Figure 4. *The first two dimensions of the Principal Component Analysis of the acoustic features of "eh". In the horizontal dimension we find fundamental frequency to be dominant, and in the vertical dimension, we find voice*

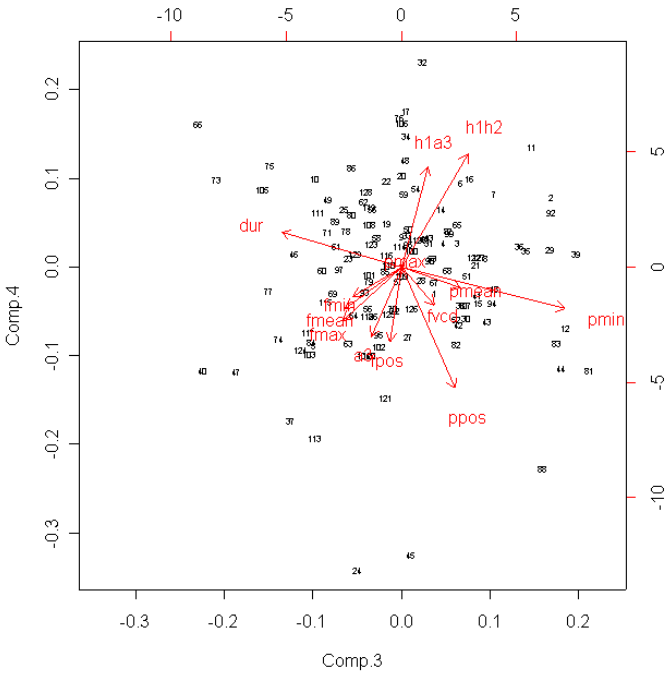

Figure 5. *The third and fourth dimensions of the Principal Component Analysis of the acoustic features of "eh". Here, the horizontal dimension appears to represent force of utterance, and the vertical dimension loudness*

6. Discussion

This paper has described an analysis of the acoustic features of a set of 129 utterances of the interjection "eh" in conversational speech. We have shown elsewhere that non-native listeners from different cultural backgrounds were able to correctly perceive the discourse and affective information in the majority of the utterances, and here described the acoustic features that best explain their distinctive characteristics by means of a principal component analysis.

We found that in addition to the conventional prosodic parameters of pitch, power, and duration, voice-quality emerges as an important discriminator. The tone-of-voice of an utterance is known to be of perceptual significance to human listeners, and its acoustic correlates of breathiness or increased spectral tilt can be measured by several methods. Here, we have seen that the *h1a3* measure proposed by Hanson and Sluijter is an effective discriminator that can be easily calculated by means of a simple tcl function. It appears that in addition to the pitch of each utterance, the power of the voice is also important in discriminating between different uses of the interjection "eh", which can express a variety of meanings or pragmatic effects in conversation. From the pca of the Japanese listeners data we determined that at least 15 different 'meanings' can be discriminated reliably.

7. Conclusion

This work is part of the ATR/JST Expressive Speech Processing project for the design of human-friendly speech technology. It shows not only that non-verbal utterances play an important part in speech perception, but also that the technology is available for processing such information for the machine understanding of human communication. We now aim to extend this work to other and more varied speech utterances, and eventually incorporate the results in both speech recognition and synthesis.

8. Acknowledgements

The author is grateful to the JST-CREST for enabling the collection of the data, to Donna Erickson for collecting the perceptual labels, and to members of the ESP project for help with the production and analysis of the data.. This work was partly supported by the Telecommunications Advancement Organisation of Japan.

9. References

[1] Campbell, N., 2002. Recording techniques for capturing natural everyday speech. *Proc LREC*.
[2] Campbell, N., 2004. The JST Expressive Speech Corpus. *Proc LREC*.
[3] Campbell & Mokhtarri, 2003. Voice Quality, the 4th prosodic dimension. *Proc ICPhS*, Barcelona.
[4] Campbell, N., 2003. Voice characteristics of spontaneous speech. *Proc ASJ Spring meeting*.
[5] Campbell, N., Erickson, D., 2004. What do people hear? A study of the perception of non-verbal affective information in conversational speech. *Journal of the Phonetic Society of Japan*, Vol.7, n°4.
[6] Schlosberg, H., 1952. The description of facial emotion in terms of two dimensions. *Journal of Experimenal Psychology*, 44: 229-237.
[7] ESP pages: http://www.feast.his.atr.co.jp
[8] Snack : http://www.speech.kth.se/snack
[9] Hanson, H., 1967. Unpublished PhD thesis, MIT.
[10] Sluijter, A., et al, 1997. Spectral balance as a cue in the perception of linguistic stress. Journal of the Acoustical Society of America, 101(1): 503-513.
[11] Campbell, N., Beckman, M., 1997. Stress, prominence, and spectral tilt. In Botinis et al (eds) *Intonation: analysis, modelling, and technology*, ESCA.
[12] Mokhtari, P., Campbell, N., 2002. Automatic characterisation of quasi-syllabic units for speech synthesis based on acoustic parameter trajectories: a proposal and first results. *Proc. Autumn meet.ing of the Acoust. Soc. Japan*, Akita: 233-234.
[13] Ihaka, R., Gentleman, R., 1996. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, 5, 3: 299-314.