Perception of Affect in Speech - towards an Automatic Processing of Paralinguistic Information in Spoken Conversation

Nick Campbell

ATR Network Informatics Laboratory
Keihanna Science City, Kyoto 619-0288, Japan

nick@atr.jp

Abstract

This paper reports a study of the perception of affective information in conversational speech utterances and shows that there are consistent differences in the acoustic features of same-word utterances that are perceived as having different discourse effects or displaying different affective states. We propose that rather than selecting one label to describe each utterance, a vector of activations across a range of features may be more appropriate. This finding complicates the representation of speech elements, but offers a more appropriate description of their attributes.

1. Introduction

People display affect in many ways; in speech, changes in speaking style, tone-of-voice, and intonation are commonly used to express personal feelings, often at the same time as imparting information. This paper describes a study of the information carried by a spoken utterance in the context of affect perception, with the goal of processing such paralinguistic information for computer speech understanding.

It has been proposed in previous work [1, 2, 3] that speech utterances can be categorised into two main types for the purpose of automatic analysis; I-type, which are primarily information-bearing, and A-type, which serve primarily for the expression of affect. The former can be well characterised by the text of their transcription alone, but the latter tend to be much more ambiguous, and require a knowledge of their prosody before an interpretation of their meaning can be made.

In previous work looking at the utterance “Eh”, [1, 4] we have found that listeners are consistent in assigning affective and discourse-functional labels to interjections heard in isolation without contextual discourse information. Although there was some discrepancy in the exact labels selected, there was considerable agreement in the dimensions of perception (as determined by principal component analysis to be aligned on the valency-activation axes described in the psychological literature, e.g., [5]) This ability seems to be also language- and culture-independent as Korean and American listeners were largely consistent in attributing ‘meanings’ to the same Japanese utterances.

In this paper, we look in detail at an example of one such utterance; the word “honma” in Kansai Japanese (equivalent to “really” in British English), which was used by one speaker 315 times in telephone conversations with various interlocutors over a period of ten weeks. The data are part of the ESP corpus [6] which was collected as part of the Expressive Speech Processing project of the Japan Science & Technology Agency [7].

2. Functional Ambiguity of Speech

The same word can be used in different contexts to express quite different meanings. Like the word “really”, “honma” can be used as a modifier (really hot, really interesting) or as an exclamation (really!), or a question (really?), or just as a backchannel utterance to indicate that the listener is interested in a conversation and encourage the speaker to continue talking. We were interested to know whether the word is consistently pronounced differently when used in each of these situations, or whether the difference in meaning arises simply as a consequence of its situation in a dialogue. If the former, then machine processing of intended meaning from acoustic parameters should be possible.

The ESP corpus contains 1000 hours of natural conversational speech collected from a small number of speakers over a long period of time, using high-quality recording equipment in a variety of daily-life situations. We have transcribed a large part of it (about 70%) and are annotating a smaller part (about 30%) for discourse and speaking style features, including the labelling of emotional and affective information. Our labellers have considerable freedom in their choice of labels (see [3] for a list) but we are still unsure about the optimal way to categorise affect or intended meaning when labelling speech. Different listeners perceive different aspects of this multi-faceted phenomenon and it can be difficult to achieve a consensus on the choice of a single most appropriate label for any given speech utterance.

3. A Perception Experiment

In order to obtain a majority opinion for a small number of tokens, we performed a perception experiment and trained statistical models to predict similar results from acoustic features derived from the speech tokens.

We excised 315 utterance tokens from high-quality recordings of conversational speech and presented them to a group of 24 listeners, divided into three subgroups, who heard 105 tokens each and noted their characteristics by means of an interactive web-page. We encouraged our subjects to listen to the speech samples using headphones, but the equipment was not standardised and no controls were exerted over the listening conditions. The listeners were asked to note not just the ‘emotion’ they perceived, but also the ‘function’ of each utterance by clicking buttons on the page after listening to the speech sound as many times as they considered necessary. They were offered the choices listed in Table 1, which were determined after preliminary trials with open-choice responses for the same data. The descriptors labelling affective states (‘perception’) were mapped to radio-buttons so that only one label could be selected, but those indicating discourse function allowed multiple responses by use of check-buttons.

Table 1: Perception of affect and functional categories used in the listening test for the utterance "Honma". Numbers count the times a given label was selected overall. A further category ('pass', n=51) was offered to allow listeners to skip a token. The category 'aizuchi' refers to a type of (nodding) back-channel response which is very common in Japanese

Perception:		Function:	
disappointed	226	aizuchi	981
disgusted	180	adjective	57
doubtful	380	laugh	227
happy	387	ok	415
impressed	412	other	65
satisfied	298	very	125
surprised	406	question	471
unhappy	227	understanding	648
	2516		2989

We can see from Table 2 that only a few listeners opted to use multiple responses. One enthusiastic listener (C-204) responded to almost all tokens of every group, but several listeners failed to respond to the complete set of all tokens. The no-answer category ('pass') was used 51 times. Subjects were university graduate engineering students and were rewarded with a grade-point for their cooperation in the experiment.

Table 3 details the responses by group. The remarkable similarity in distribution of responses confirms that the data can be considered sufficiently representative and that, in general, all three groups responded to the sounds in a very similar way. However, the number of exact matches across the responses was very small and considerable individual differences were noted. The following samples illustrate the types of 'disagreement' found. Less than 10% of responses were unanimous, but we can see from the table that in most cases the reponses are complimentary rather than contradictory. Utterance no.18, for example, is rated positively in all but one case. Utterance no.36 appears to be more negative, although one listener perceived satisfaction instead. Utterance no.64 was largely perceived as happy, but the only other response was exactly the opposite.

Sample response counts from the listeners:

utt 18: happy 1, impressed 1, satisfied 2, surprised 1, unhappy 1
utt 19: doubtful 1, happy 1, impressed 2, satisfied 1, surprised 2
utt 31: doubtful 1, happy 5, impressed 1
utt 32: doubtful 2, happy 3, satisfied 2
utt 33: doubtful 1, surprised 6
utt 36: disappointed 3, disgusted 1, doubtful 2, satisfied 1
utt 37: doubtful 3, happy 4
utt 38: disgusted 2, doubtful 2, impressed 1, unhappy 1
utt 64: happy 5, unhappy 1

utt 18: aizuchi 3, adj 1, laugh 1, ok 1,understanding 4
utt 19: aizuchi 2, laugh 1, ok 1, question 2, understanding 2, very 1
utt 31: aizuchi 2, laugh 4, ok 2, understanding 1, very 1
utt 32: aizuchi 3, laugh 3, ok 1, question 1, understanding 2
utt 33: aizuchi 1, adj 1, ok 2, question 2, understanding 2
utt 36: aizuchi 3, ok 2, question 2, understanding 1
utt 37: aizuchi 1, adj 1, laugh 3, ok 1, question 2
utt 38: laugh 1, ok 1, question 2, understanding 2
utt 64: laugh 4, other1, very 1

Table 2: Counts of responses by category

group	lstnr-id	percep	func
A	030	103	135
A	043	105	105
A	065	105	105
A	069	105	143
A	106	104	104
A	120	103	162
A	205	102	148
B	020	83	83
B	028	105	118
B	031	105	129
B	059	105	105
B	077	105	105
B	101	105	105
B	119	50	50
B	126	101	197
B	138	98	114
C	056	99	117
C	037	103	155
C	046	104	147
C	047	103	128
C	088	105	123
C	111	68	106
C	145	99	105
C	204	251	251

It is not unexpected that listeners should perceive an utterance differently, particularly when stripped of discourse context information, but it is of interest to know whether these results indicate that they are simply perceiving different aspects or whether they are using different descriptors for what may be basically the same 'colouration' of the speech sounds. To resolve this problem, we constructed a statistical model of the results and found that there is an underlying consistency in the patterns of responses.

4. Statistical Modelling of the Results

From a manual analysis of the responses, we built a list of optimal consensus labels for the speech tokens from the majority responses, introducing compound categories (such as d-q : disappointed question or i-u : impressed understanding) to resolve some of the disparity in labels. We also introduced a 'mixed' category for those tokens where listeners just didn't agree at all on an appropriate label. Finally, any categories that had fewer than five tokens were grouped into a garbage category using the label 'xxx'. This resulted in a set of 45 backchannels, 16 disappointed questions, 15 happy, 7 laughing, 15 happy-and-laughing, 22 surprised, 25 understanding, 15 okay, 20 question, 16 impressed, etc., with 86 mixed-label tokens.

4.1. Training from Acoustic Parameters

We then built a classification tree (using the public-domain 'R' statistical software package [9, 10]) to learn the relationships between the acoustics and the perceptual characteristics in order to predict the most likely response for each speech token for a reclassification. We used simple first-order statistics derived

Table 3: Different data were presented to the three groups of listeners, but the distributions of the responses appear to be very similar

	A	B	C
disappointed	68	65	93
disgusted	35	86	59
doubtful	114	120	146
happy	128	125	134
impressed	127	135	150
satisfied	101	97	100
surprised	117	147	142
unhappy	37	82	108
	A	B	C
aizuchi	263	408	310
adjective	14	8	35
laugh	98	46	83
ok	122	117	176
other	3	26	36
very behind	25	29	71
question	138	137	196
understanding	228	214	206

from the acoustics as the independent variables.

Previous work had confirmed the following features to be useful: utterance duration, f0-range, f0-variation, f0-maximum, f0-minimum, f0-mean, position of f0 peak in the utterance, position of f0-minimum in the utterance, power-range, power-variation, power-maximum, power-minimum, power-mean, position of power-peak in the utterance, position of the power-minimum in the utterance.

The tree correctly predicted 68% (or 217/315) of categories using 26 leaf nodes. The resulting predictions and details of the reclassification of the tokens based on the trained tree are available at http://feast.his.atr.jp/data/honma/pred.html, where the speech tokens can be listened to interactively and the reader can assess the results personally. Since it appears that the tree is able to generalise on the basis of common acoustic characteristics and to produce an appropriate general classification, we performed a manual analysis of the supposed errors (and successes) in order to determine whether they were perhaps improvements in the classification rather than misclassifications. The materials for this analysis can be found at the above site.

Of course, the prediction tree matches well with the majority of listener responses, but we do not believe that this in itself is a satisfactory conclusion. It does not explain the differences in opinion of the individual respondents when they listened to the same or equivalent samples. However, underlying the single-right-answer that is output by a classification tree is a vector of probabilities for each possible response in the category space. Closer examination of these probability vectors revealed that although the final decision was made according to the class having the highest probability, there were many cases when the next-best, or even the top-three, classes had smaller but very similar probabilities. We therefore changed our approach, and instead of building one overall classifier, we decided to build a parallel set of classifiers, each outputting a probability for its own category.

Table 4: Sample predictions from the tree trained on acoustic features. Numbers after the categories represent percent activation of each of the features. Activation levels lower than 20% are not considered to be relevant. Speech samples can be accessed at the web page noted above

18: understanding 40 aiduchi 33 very 27 satisfied 25 impressed 23
19: satisfied 31 happy 28
31: happy 67 laugh 44 aiduchi 40
32: happy 67 laugh 66 aiduchi 55
33: surprised 56 understanding 34 aiduchi 33 question 20
36: aiduchi 54 question 20
37: doubtful 29 happy 25 laugh 23
38: disgusted 33 impressed 32 understanding 25 unhappy 23
47: question 58
64: happy 67 laugh 44
196: aiduchi 41 laugh 38

4.2. Training a Set of Classifiers

When checking the predictions of the tree classifier trained on the optimised labels, we were not just interested in the true-or-false one-right-answer, but also in explaining the multiplicity of responses from the human listeners. We therefore used the predict() function of the 'R' software to produce a matrix of output probabilities rather than a single-correct-answer. By listening to each waveform sample and observing the probabilities modelled by the classifier, it became apparent that multiple responses were appropriate.

We therefore built a table of likely responses for each speech token by summing (and normalising) the individual responses from all listeners for each category. This resulted in a matrix of 'probabilities' for every category for each token which we used as the dependent variables for a further set of classification trees. This time, however, rather than grow a single tree to predict a single category response for each token, we grew as many trees as there were categories in the original data from the human listeners. This resulted in a set of sisteen trees. Each tree was trained with the same acoustic data for the independent variables, and with the probabilities (or normalised counts of human responses) of a response in its own category as the dependent variable.

After training, we used a function that passed the same acoustic data to each of the classification trees, and allowed each to output a probability indicating the likelihood of a human listener entering a response of its own category. These likelihoods were ranked and an ordered list of relevant categories was produced for each speech token. These were then thresholded, using a limit determined experimentally, so that a vector of most likely responses for each probable class was produced. Table 4 illustrates type of output produced by the classifier. For ease of reading, the 'likelihoods' of each category are printed as a percentage figure. The full table of results can be accessed at http://feast.his.atr.jp/data/honma/pred2.html.

Evaluation of this method of classification is no longer easy, as simple counts are no longer appropriate, but perceptual evaluations (informal listening tests) have been very encouraging. The method is similar to that proposed by Wightman [11] for the determination of an appropriate prosodic boundary score in ToBI label analysis; using counts of actual human responses to represent likelihoods of an actual boundary.

5. Discussion

This paper is not so much concerned with the degree of success of the different statistical classifiers; rather it is more about how we categorise speech. When producing a database of labels to annotate a corpus, we fall too easily into the trap of thinking that there is one right answer in any given situation. For syntactic analysis, this might be the case; a noun is a noun, an adjective is an adjective, and even if the same lexical word might have a potentially ambiguous syntactic status, it is usually very clear from the context which particular syntactic role a word is taking in any given situation. Ambiguity of classification is not considered.

However, for the labelling of affective information in speech, we find that different people are sensitive to different facets of information, and that e.g., a question may function as a backchannel, and a laugh may show surprise at the same time as revealing that the speaker is also happy. A happy person may be speaking of a sad event and elements of both (apparently contradictory) emotions may be present in the speech at the same time.

This work has helped us to understand the complexity of affect in a spoken utterance. As a result, we believe that there is no one-right-answer where the labelling of affect is concerned, and that it is better to represent this type of paralinguistic information by using a vector of probabilities instead. That is, a given utterance may evoke response ‘a’ from ‘x’ percent of the listeners, and response ‘b’ from ‘y’ percent, and response ‘c’ from ‘z’ percent, etc.

In a computer speech understanding system, this would be best represented as a set of daemons each tuned to one aspect of speech, and their collective activation or energy used as an indicator of the affective colouring of each utterance. Not one descriptor, but an emergent field of activations to represent the state of the speaker and the nature of the speaker-listener relations as well as the intended function or purpose of the utterance.

This is not a simple view, but then speech is a very complicated information source.

6. Conclusions

This paper has described a perception experiment that was designed to collect data on the presence of affective information in conversational speech samples, and particularly on the extent to which this informaion can be recognised without any discourse context information being available. Materials were collected by use of interactive web pages from a small number of listeners, and tree-classifiers were trained to predict category labels on the basis of these results.

We found that there was considerable variation in the responses from the listeners and that in very few cases was the exact same label selected for any given speech utterance. However, we conclude that rather than indicating a disagreement between listeners, this indicates that more than one label may be necessary to describe any given speech utterance. Different listeners may perceive different facets of an utterance, so we propose that rather than selecting any one label as optimal in any given case, it may be more appropriate to use a vector of activations across a range of features in every case. This finding complicates the representation of speech elements, but offers a more appropriate description of their attributes.

7. Acknowledgements

This work is partly supported by the Japan Science & Technology Corporation (JST), and partly by the National Institute of Information and Communications Technology (NICT). The author is grateful to the management of ATR for their continuing encouragement and support, and to the students of NAIST for their cooperation with the experiments.

8. References

[1] N. Campbell, (2004) “Listening between the lines; a study of paralinguistic information carried by tone-of-voice”, pp 13-16, in Proc Internbational Symposium on Tonal Aspects of Languages, TAL2004, Beijing, China.

[2] N. Campbell, (2004) ”Getting to the heart of the matter”, Keynote speech in Proc Language Resources and Evaluation Conference (LREC-04), Lisbon, Portugal.

[3] N. Campbell, (2004) “Extra-Semantic Protocols; Input Requirements for the Synthesis of Dialogue Speech” in *Affective Dialogue Systems*, Eds Andre, E., Dybkjaer, L., Minker, W., & Heisterkamp, P., Springer Verlag.

[4] N. Campbell & D. Erickson, (2004) ”What do people hear? A study of the perception of non-verbal affective information in conversational speech”, in Journal of the Phonetic Society of Japan, V7,N4.

[5] Schlosberg, H., (1952) “The description of facial emotion in terms of two dimensions”, Journal of Experimenal Psychology, 44,229-237.

[6] N. Campbell, (2002) ”Recording Techniques for capturing natural everyday speech” pp.2029-2032, in Proc Language Resources and Evaluation Conference (LREC-02), Las Palmas, Spain.

[7] The Japan Science & Technology Agency *Core Research for Evolutional Science & Technology*, 2000-2005

[8] N. Campbell & P. Mokhtari (2003) “Voice quality: the 4th prosodic dimension”, in pp.2417-2420 in Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS’03), Barcelona, Spain.

[9] R. Ihaka, and R. Gentleman, (1996) “R: A Language for Data Analysis and Graphics”, Journal of Computational and Graphical Statistics, vol5.3,pp.299-314.

[10] R-project web pages, (2004) the Comprehensive R Archive Network (www.r-project.org)

[11] C. W. Wightman and R. C. Rose, (1999) “Evaluation of an Efficient Prosody Labeling System for Spontaneous Speech Utterances”, . In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, Keystone, CO.