What do people hear? A study of the perception of non-verbal affective information in conversational speech.

Nick Campbell[1] & Donna Erickson[2]

[1]Research Director, JST/CREST ESP Project, ATR, Kyoto Japan

[2]Professor, Gifu City Women's College, Gifu, Japan

Abstract (99 words)

This paper examines perceptual characteristics of non-verbal speech, focusing on "eh", defined in the dictionary as "interjection" or "yes", but used in colloquial speech to invoke a number of different conversational effects. Perception tests with speech taken from the Expressive Speech Corpus were given to Japanese, Korean, and American English listeners who were asked to "sort" the data into different "boxes" on the computer screen. PCA analysis showed that although listeners were not unananimous in their selection of labels, they tended to group the utterances according to their affective valences (positive or negative affect) and strengths (passive or active).

Key words: perception, affective information, cross-linguistic, colloquial speech

Introduction

This paper examines the perceptual characteristics of non-verbal speech in human communication, with a focus on prosodic variety and corresponding listeners' perceptions. From a very large corpus of conversational spoken Japanese, we selected one of the most common speech tokens and asked listeners to indicate the perceived affect in different renditions of a given utterance from the same speaker.

The term "affect" or "affective information" is used here to refer to paralinguistic information that is present in spoken utterances as part of the speaker's message to the listener. This affective information is not coded semantically in the dictionary definitions of a word, but is highly context-dependent, depending on the situation of the conversation as it occurs at the time, e.g., who the speaker is (i.e., what aspect(s) of their persona they are presenting), who they are talking to, where, why, formality, familiarity, etc. Changes in prosody, i.e., duration, pitch, loudness, tone of voice (voice quality), are often used as the vehicle used for conveying such affective information [1]. A growing number of studies are available in the literature about the importance of prosodic

changes in a word or sentence for conveying information to the listener about the speaker's mental or emotional attitude (e.g., [2]), as well as about the relationship between the speaker and listener, i.e., friend, same age or social class, parent, spouse, lover, etc. [3], or the context in which the utterances are spoken, i.e., formal/informal meeting, telephone conversation to a business/friend, etc. [4].

The utterances selected for the experiment all consisted of the single monosyllabic word "eh," which is used in colloquial speech in a variety of contexts to invoke a number of different conversational effects, but primarily to show affective information. Although phonetically the tokens we studied were extremely simple, they were prosodically very rich and varied in terms of length, F0 pattern, loudness, voice quality, etc. Our principal interest was to determine the extent to which different listeners perceived the same affect from different utterances having similar prosodic characteristics; or perceived different affects from utterances with the same "segmental" characteristics, but having different prosodic characteristics. The interested reader can access and listen to the original speech files and see their associated plots and data at the JST/CREST ESP project web-pages [5].

According to the literature, the term "eh" can be used in Japanese either as a backchannel, signifying mild agreement to indicate simply that the listener is paying attention to the speaker, or it can be used in its strong form to mean lexical "yes." We were initially interested in determining the prosodic characteristics that distinguish these two forms, for use in speech technology applications, but soon realised that the perceived meanings of such an utterance are far more complicated than we had at first thought. The simple interjection "eh" functions to express a variety of speaker-listener and speaker-state relationships, and listeners appear to be sensitive to these at many different levels and combinations. All conversations involve interaction between two or more people, and as such, the affective meaning of a single utterance within a conversation/speech event is highly context-dependent—who is speaking and to whom; what is their relationship (to each other and to themselves), and where/why are they speaking/listening? The context-dependency, of course, helps the listener-speaker to communicate.

However, a first step to isolate some of the various characteristics that signal affective information in a spoken-exchange is to ask listeners to identify the affect of an utterance, independent of context, i.e., to present a single word to listeners and ask

them to say what is the affect. Previous studies [e.g. 6,7] have reported that listeners are able to label affective information with a certain degree of reliability, just by listening to isolated utterances. The purpose of this study is to explore this more intensively, using a large sampling of natural speech. The data used for the experiment reported here consisted of 129 instances of "eh" spoken by a mature female speaker of standard Japanese in a variety of conversational situations over a period of three months to several interlocutors who differed in sex, familiarity, and age. The recordings were made in an acoustically damped environment using head-mounted studio-quality microphones. All conversations were held using a telephone line, with the interlocutor at a distant location, so no visual information was available. High quality recordings were taken to DAT tape while the speaker held a series of informal 30-minute conversations over the telephone once a week. Since the utterance "eh" itself is extremely simple, yet the information it can carry is complex, we assume that the manner of speaking, or speaking style, can be the only medium or carrier of this communicative information. Previous work has reported on the effect of the interlocutor on differences in speaking-style parameters [4], but the present work tests the extent to which different listeners can perceive the same intended communicative effect even when no specific information regarding utterance context is available; i.e., we measured the information carried only in the physical utterance itself. Utterance tokens were excised from recordings of the fluent conversational speech and were presented to listeners using the purpose-designed computer software interface [8] described below. Although our original intention was to determine the extent to which native listeners agreed in their perceptions of the intended function(s) of an utterance, we were also interested in knowing the extent to which listeners who were not speakers of the same language were able to perceive the same or similar effects. Previous studies have shown that native language affects perception of affect [9,10]. We therefore carried out an initial perception test using 21 native speakers of Japanese, and then performed similar (but somewhat restricted) tests with the help of 12 native speakers of Korean in Korea and 19 native speakers of American English in the United States of America. None of the non-Japanese respondents were familiar with the Japanese language.

Perceptual Data Collection

Speech samples were taken from the Expressive Speech Corpus [5], which has been described extensively elsewhere [11,12]. Since “affect” tends to be perceived

unconsciously, it is not a simple matter to notate or describe a listeners' perception of the speaker's affect. Our approach to this problem was to use a software interface written in the TCL/TK programming language to allow subjects to listen to individual speech tokens and to categorise them freely without time or space restrictions.

The initial state of the software interface is shown in figure 1, left part, with movable circles representing the speech samples aligned in random order along the main diagonal, and the rest of the screen blank. In a way similar to sorting socks, books, or gramophone records, listeners were first required to determine from the data the types and number of categories that they considered appropriate, and then to determine which specific items belonged to each category. These two forms of decision were usually made concurrently throughout the sorting process.

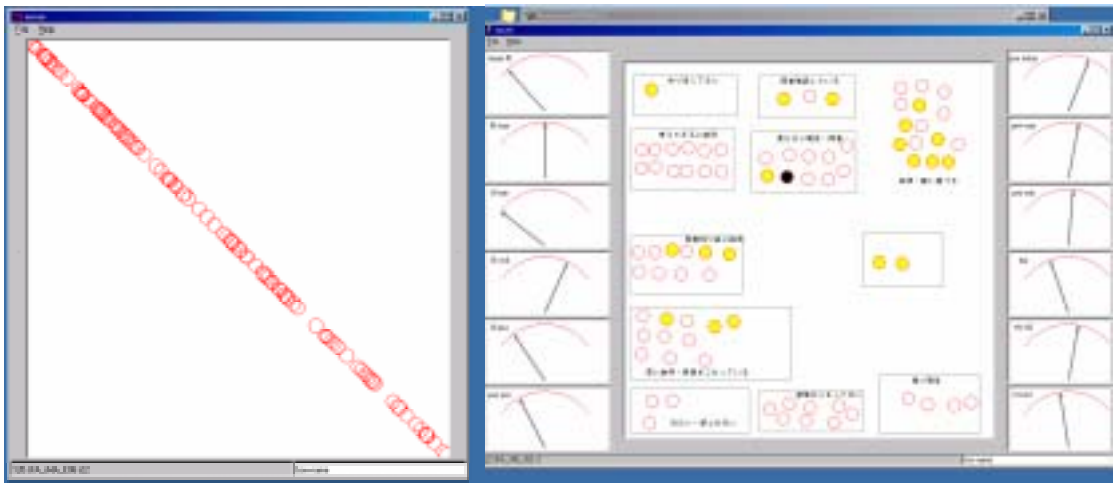

Figure 1. The left-hand panel of the figure shows a screen-shot of the initial position, before labelling has started. The right-hand panel shows the final state, with all (or most) tokens in labelled boxes. In this example, no boxes are nested or overlapped. The meters on the left and right edges of the right-hand panel are not normally displayed during the labelling process but are included in the figure to show how the acoustic characteristics of each token can be checked visually by researchers after the labelling task has been completed. Although acoustic analysis was not included as part of this current study, it is part of the ongoing aspect of this research [13].

No restrictions were placed on the number or types of categories other than those implied by the physical dimensions of the screen. Circles (representing individual speech tokens) can be listened to and moved anywhere within the screen by clicking or

dragging with the computer mouse. A group of circles can then be named by drawing a box around them, and labelling the box with a text string using ASCII characters. Categories (i.e., boxes) can be overlapping or nested if required.

The middle part of the right-hand panel of Figure 1 shows a typical screen after sorting has taken place. Samples have been placed in boxes that have labels decided by the subject. The number of boxes, and the names assigned to them, are freely determined by each participant. No guidelines or assistance, other than with the basic functionality of the software, was provided. In this way, we were able to compare not only the agreement between subjects with respect to the classification of individual tokens, but also with respect to the perceived categories themselves and the types of descriptors used. Respondents typically required about forty minutes to complete the task, which consisted of listening to and classifying each of the 129 short monosyllabic utterances.

A second computer program then processed the log files produced by the classification interface software and produced (a) a list of what was listened to when by who, and (b) a set of labels for each data point. The listening log provides a revealing measure of task difficulty by noting the number of times a given utterance was listened to and at which position on the screen, but for reasons of space, only the labelling results will be reported in the present paper. Data from all listeners were collected and analysed using the free public-domain "R" statistical analysis software from [14].

Tables 1 and 2 show sample responses from the set of English-speaking listeners. It is clear from these tables that the labelling was by no means unanimous. However, it is also clear that certain utterances were well identified. The tables show conflicting labels (*happy* and *sad*), and e.g, that utterance No.30 and No.108 were distinctively *happy*-sounding, whereas utterance No.66 was perceived as *sad*. However, one listener labelled No.66 as *happy*. Some utterances (e.g., No.10) have the same number of responses for both *happy* and *sad*. This indicates a difference of opinion between the labellers. The aim of this paper is to determine the extent to which different listeners heard the same effects in order to test the non-verbal perception of speech.

Table 1. Number of responses from all English listeners for "happy." The table lists all speech samples in three rows, showing the utterance id number (id) and the number of "happy" responses for that utterance (n=). Compare these responses with the "sad" responses below.

id	1	6	8	9	10	12	13	15	17	23	26	28	29	30	31	33	35	36	39	43	44	45	50

n=	1	2	7	2	3	2	3	1	2	1	1	9	1	11	1	3	2	1	1	1	4	2	2	
id	51	52	54	55	59	60	61	63	64	66	68	69	70	73	74	76	77	78	80	81	82	83	84	
n=	5	7	6	2	2	1	1	1	5	1	3	1	2	2	4	1	4	2	1	1	5	2	6	
id	85	87	88	90	95	96	97	101	103	104	105	106	108	112	113	114	116	117	122	123	125	128		
n=	1	1	1	1	3	2	2	2	3	1	1	1	12	1	1	3	3	1	2	8	3	5		

Table 2. Number of responses from all English listeners for "sad." The table lists all speech samples in three rows, showing the utterance id number (id) and the number of "sad" responses for that utterance (n=). Compare these responses with the "happy" responses above.

id	3	5	9	10	11	12	14	19	21	22	23	24	25	26	27	29	31	32	34	35	36	39	41	
n=	1	3	2	3	3	1	1	1	2	1	3	1	2	1	1	2	1	1	3	3	2	3	1	
id	42	43	45	46	48	49	50	56	59	60	61	63	66	72	73	74	76	77	82	84	86	87	90	
n=	1	4	3	1	5	2	2	1	2	1	5	2	6	1	2	2	4	1	1	1	3	1	2	
id	91	94	95	96	100	103	106	109	110	113	119	121	123	124	126	127	128							
n=	2	3	1	1	2	2	1	1	1	5	1	3	1	3	5	5	3							

Table 3. Confusion matrix for all English listeners, with percentage of total responses. Reading down the table, counts are given (along the rows) for all categories of response for each label; e.g., of the set of utterances that were assigned the label "ead" (n=135) by any listener, some were also identified by other listeners as "ean" (n=53), none as "eco", some as "eex" (n=24), some as "eha" (n=65) etc., where n represents the total number of responses across all listeners and tokens combined. We would expect to find the largest numbers along the main diagonal; although this is the general tendency, it is not always the case. The meaning of the labels in the left-hand column, i.e., "ead" is indicated by the labels in the left-hand column, i.e., "annoyed."

	ead	ean	eco	eex	eha	ehe	eht	ein	eiu	erl	esc	esd	eso	esp	eun	eus	%	label
ead	135	53	0	24	65	6	50	70	50	51	65	45	73	18	15	50	9.1	annoyed
ean	74	94	0	28	33	2	48	26	20	23	23	48	63	16	7	39	6.3	angry
eco	0	0	12	21	23	1	0	2	2	1	0	2	0	21	1	4	0.8	questioning
eex	27	28	8	122	110	8	36	18	18	28	35	70	23	62	5	25	8.2	excited
eha	33	18	10	99	186	12	36	55	40	39	49	74	13	70	7	34	12.5	happy
ehe	9	2	1	18	21	14	3	15	12	8	0	9	3	5	1	6	0.9	hello
eht	66	64	0	40	49	2	92	37	29	31	75	82	57	15	8	33	6.2	hurt
ein	72	27	2	30	82	9	28	108	63	47	38	34	40	20	19	54	7.2	indifferent
eiu	51	22	2	24	57	8	21	74	82	29	21	40	25	22	14	43	5.5	indiff/uncaring

erl	60	36	1	26	50	6	38	44	37	83	48	57	33	17	9	32	-	5.6 relieved
esc	57	17	0	23	72	0	39	31	27	34	116	61	46	6	14	29	-	7.8 sad-crying
esd	48	39	1	69	73	6	58	40	36	44	86	135	43	31	12	60	-	9.1 scared
eso	91	69	0	21	22	2	55	29	21	33	57	47	95	8	14	41	-	6.4 sick-of
esp	20	18	9	89	87	5	19	26	21	23	7	50	9	82	4	18	-	5.5 surprised
eun	34	13	1	12	26	1	8	33	21	13	23	18	27	10	25	26	-	1.7 unconcerned
eus	72	33	4	40	59	6	32	78	58	38	44	56	40	30	16	110	-	7.4 unsure

Table 4. Confusion matrix for all Korean listeners. Reading down the table, counts are given (along the rows) for all categories of response for each label; e.g., of the set of utterances that were given the label "kap" (n=78), some were also identified as "kas" (n=24), some as "kay" (n=5), some as "kch" (n=10) etc., where n represents the total number of responses across all listeners and tokens combined. We would expect to find the largest numbers along the main diagonal, and here this is indeed the case, indicating more conformity or agreement among the Koreans than among the American listeners for these Japanese "eh" utterances.

	kap	kas	kay	kch	kda	kgc	khb	khd	khn	km2	km3	kmu	knl	kpl	ksp	kta	label(translation)
kap	78	24	5	10	22	42	18	6	21	3	25	14	4	16	31	24	apuda (sick-of)
kas	22	69	17	6	29	23	17	15	14	5	11	20	13	15	16	19	ansimhada (relieved)
kay	6	27	30	2	16	10	16	8	12	3	7	12	15	10	4	10	anneyong (hello)
kch	16	12	2	15	2	18	2	3	3	2	5	5	1	3	12	4	chilmunhada (confused)
kda	22	32	15	2	79	25	15	18	22	4	18	20	21	18	16	33	darum (other)
kgc	44	23	7	11	30	71	7	4	21	4	15	18	6	12	21	21	guichanta (annoyed)
khb	18	23	12	2	21	7	50	14	8	1	8	10	22	12	14	15	haengbokhada (happy)
khd	10	19	5	3	22	5	17	45	15	1	19	8	29	12	13	21	haengbunhada (excited)
khn	20	16	10	3	22	20	10	13	48	0	16	6	15	8	13	19	hwaganada (angry)
km2	4	7	3	2	7	9	1	1	0	10	1	11	0	0	3	2	musimhan (indiff/unc)
km3	20	12	6	5	16	17	9	12	22	1	50	12	11	13	11	19	musupda (scared)
kmu	14	26	11	3	24	22	10	8	9	9	14	42	8	9	10	13	musimhada (indifferent)
knl	4	11	10	1	20	6	20	26	11	0	14	6	64	17	6	14	nolada (surprised)
kpl	25	17	11	3	22	15	15	11	10	0	15	8	26	46	13	16	bulanjuhghan (unsure)
ksp	35	23	4	9	21	29	15	9	12	3	14	13	4	11	55	17	sulpuda (sad/crying)
kta	23	23	11	4	31	19	14	13	16	2	13	14	14	12	21	61	dachida (hurt)

Table 5. Confusion matrix for all Japanese listeners. Reading down the table, counts are

given (along the rows) for all categories of response for each label; e.g., of the set of utterances that were given the label "jai" (n=27), some were also identified as "jbi" (n=15), some as "jdo" (n=51), some as "jfm" (n=16) etc., where n represents the total number of responses across speakers and tokens. We would expect to find the largest numbers along the main diagonal, but this is not seen here.

	jai	jbi	jdo	jfm	jgk	jgm	jhm	jht	jiy	jk2	jki	jkn	jko	jks	jkz	jnb	jnm	jnt	jny	jod	jok	jsb	jto	jun	jut	jya	jyo	jzb
jai	27	15	51	16	1	17	4	12	13	1	59	16	8	20	4	17	5	28	11	64	7	3	11	21	7	10	1	
jbi	13	65	10	20	12	44	22	21	27	18	108	32	40	1	5	6	8	0	7	328	18	7	34	1	6	7	32	4
jdo	10	8	70	19	9	2	8	14	24	14	4	25	16	17	3	16	9	26	6	60	18	3	11	19	4	4	1	3
jfm	13	18	49	77	12	25	14	27	115	9	46	46	68	25	17	83	14	29	26	156	26	14	31	24	14	11	14	2
jgk	1	13	11	15	30	5	7	10	39	16	17	26	35	4	4	23	1	2	7	110	12	2	19	1	1	2	21	3
jgm	9	29	3	16	4	79	5	10	9	179	0	6	16	0	1	12	4	1	10	251	3	5	23	4	12	3	50	1
jhm	4	25	10	20	8	9	31	21	34	28	27	33	41	0	6	7	5	0	2	127	20	8	33	3	0	6	4	8
jht	10	26	39	28	10	9	19	56	62	27	26	42	57	7	11	40	9	10	13	159	24	13	31	17	3	9	16	7
jiy	7	20	18	57	18	7	18	32	153	32	16	51	99	16	21	109	14	12	30	131	33	19	41	16	13	16	14	9
jk2	1	9	9	5	9	0	13	15	37	36	2	26	28	0	6	13	1	1	2	18	15	8	17	1	1	2	0	9
jki	13	35	5	21	8	72	11	11	15	4	196	4	28	0	1	15	4	0	11	274	2	2	30	4	9	2	47	2
jkn	8	17	18	31	14	7	14	23	63	24	12	76	48	12	11	46	11	9	15	63	36	15	25	8	8	7	14	7
jko	7	31	9	47	22	25	22	36	129	27	54	49	131	8	19	104	16	3	25	267	26	17	49	11	11	18	38	7
jks	9	2	35	13	3	0	7	0	17	0	0	12	6	28	2	18	6	25	5	27	5	2	3	17	5	2	2	0
jkz	4	5	9	27	4	1	6	14	74	12	1	16	45	5	22	31	5	7	12	20	8	7	12	5	6	5	2	3
jnb	6	6	16	31	4	8	6	14	82	7	12	29	48	13	11	122	11	9	26	52	17	8	15	19	11	6	3	3
jnm	5	10	14	16	2	4	5	12	27	1	5	22	32	17	4	31	25	11	11	116	13	7	7	8	5	7	4	1
jnt	9	0	47	14	2	1	0	6	12	1	0	9	3	20	4	5	4	32	1	5	2	2	1	17	4	0	2	0
jny	9	9	9	27	5	15	2	11	58	2	38	22	39	4	9	96	10	1	39	75	15	11	12	14	4	5	7	1
jod	16	60	20	50	23	77	26	42	95	19	195	45	99	8	11	24	14	6	16	567	22	10	65	6	16	16	77	5
jok	4	15	19	20	13	2	16	21	49	30	2	56	42	5	7	57	11	1	14	60	44	16	26	8	6	7	11	9
jsb	3	9	5	15	3	6	8	16	31	13	9	29	32	2	6	39	6	1	11	50	20	24	11	6	1	7	9	3
jto	10	36	12	32	16	25	26	31	74	28	69	38	69	3	11	31	7	1	12	269	24	11	72	3	6	10	35	8
jun	12	1	51	19	1	3	0	11	34	1	3	9	16	21	5	84	7	28	18	13	9	5	3	33	9	2	4	1
jut	5	6	4	15	1	12	3	4	32	2	21	12	13	11	4	16	5	10	3	62	7	1	6	6	24	3	11	2
jya	0	8	4	13	2	3	6	12	45	4	3	12	41	3	5	18	8	0	6	90	9	8	11	2	4	20	8	0
jyo	7	23	11	13	7	46	4	9	12	0	110	9	24	5	2	7	2	6	5	248	3	4	22	6	9	3	79	0
jzb	1	4	4	2	3	1	8	7	18	18	2	14	14	0	3	9	1	0	1	10	9	3	11	1	2	0	0	10

Table 6. Counts of the unique labels determined by the Japanese listeners for the utterance "eh", and the three-letter codes assigned to each for the analysis. X is an umbrella category for all the unlabelled (missed or unknown) tokens.

aizuchi	akirame	bikkuri	doui	dousiyou	fuman	gakkari
jai 16	jok 5	jbi 39	jdo 35	jnm 15	jfm 77	jgk 30
gimon	hai	hirumu	hitei	igai	ikari	itami
jgm 79	jdo 18	jhm 31	jht 10	jbi 26	jht 12	jht 15
iya	kanashii	kangaeru	kanshin	kikikaeshi	komaru	konwaku
jiy 153	jkn 76	jnb 87	jks 28	jki 196	jko 118	jto 20
koutei	kowai	kurai	kurushii	kyoufu	kyozetsu	mayoi
jdo 12	jht 7	jod 4	jk2 36	jht 8	jkz 22	jko 8
mukanshin	nani	nattoku	nayami	nemui	no-noise	nobashi
jnm 7	jai 11	jnt 32	jny 39	jnm 3	--- 2	jnb 13
norikijanai	ochikomu	odoroki	omoshiroi	oto	rakutan	reisei
jnb 16	jok 21	jod 567	jyo 1	--- 10	jok 11	jok 4
setsumei	shitsubou	shoudaku	tomadoi	tyuutyo	unazuki	uresii
jok 3	jsb 24	jod 5	jto 52	jko 5	jun 33	jyo 33
utagai	X	yabai	yorokobi	zetsubou	zikankasegi	
jut 24	--- 424	jya 20	jyo 45	jzb 10	jnb 6	

Tables 3-5 show confusion matrices for all three groups of listeners. It is immediately apparent that there is not a unanimous consensus in the choice of labels. However, whether this is because listeners perceived different effects from each utterance, or whether they simply chose similar or synonymous labels is yet to be seen. It is also obvious from Tables 3 and 5 that the majority of responses do not appear along the main diagonal, contrary to our expectations. Individual listeners differed widely in their choice of category for individual tokens. However, on closer analysis, we can see that many of the confusions are perhaps different lexical choices for the same underlying effect; for example, both *bikkuri* and *odoroki* indicate *surprise* (the later being a bit stronger), and *kikikaeshi* and *gimon* both indicate *questioning*. By allowing our respondents a free choice of terminology to express how they individually perceived each utterance, we obtain meaningful data, not just echoing our preconceived view of the speech forms, but we also face the problem of mining that data for more general concepts.

Table 5 shows the confusion matrix obtained from Japanese subjects listening to utterances from the female Japanese speaker. In all, 75 different labels were suggested by the subjects to describe their perception of the meaning or function of each utterance. There were 8 unique labels (used only once each), and 39 having less than 10 responses each. However, there were 28 labels (or groups of similar-meaning labels) that contained more than 20 listener-responses each, and these are listed in the table. For reasons of printing space, we reduced each label to a three-letter identifier (see left and top rows of each table) and show the user specified label (which was entered using Roman characters) only on the right-hand side of the table. The abbreviations in Table 5 are listed with their full label decscriptions and token-response counts in Table 6.

Since there was no attempt to balance the tokens beforehand, we must assume that those labels with only a few responses are as representative as those including several hundred responses, and that the less frequently used labels are as valid as descriptors as those which were more generally used. However, since a minimum of twenty tokens is required for statistical analysis, all labels having fewer than 20 responses each were grouped into more general categories for this analysis (e.g., “jnm” or “jok” in Table 6).

Principal Component Analysis

In order to better determine the underlying relationships revealed by these individual and seemingly idiosyncratic responses, we performed a principal component analysis of the data, using the eigenvectors of the correlation matrix to clarify the relationship between the individual components and reduce the initial complex relationship to a lower-dimensional space [15,16]. A calculation is performed to rotate the axes of the response vectors so that the principal components (those having the greatest effects) are ranked in order of importance. For the statistically naive reader, we can explain this process as similar to that of rotating a model of a human head so that the most recognisable angle can be found. A head can be viewed from many angles, but the front usually carries more information than the back or the top, and a slight shift from the full-face mid-line reveals profile information that serves to complement the full-face view, adding to the information the viewer receives, and maximising the recognisability of the face. In a similar way, we rotate the data to determine the optimal separation of the features by means of principal-component analysis.

The analysis of the Japanese listeners' data revealed that 16 components (or dimensions) would be sufficient to account for almost 90% of the data, perhaps indicating that only 16 labels would be needed. We can identify by overlaps among the labels those which might be considered redundant (or synonymous). Table 7 details the results, and shows that 24 components account for 99% of the responses. We can see that the remaining four components, contribute very little to explain the variance, and show at least four labels to be redundant. In the simplest case, two principal dimensions alone account for approximately 30% of the overall variance.

Table 7. Showing the contribution of each component in the pca analysis of the Japanese listeners' responses

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	2.1331494	1.9593256	1.62553408	1.36561249	1.33136575
Proportion of Variance	0.1625116	0.1371056	0.09437004	0.06660348	0.06330481
Cumulative Proportion	0.1625116	0.2996173	0.39398729	0.46059078	0.52389559

	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
Standard deviation	1.20364217	1.19412946	1.0794380	1.06777008	0.99147485
Proportion of Variance	0.05174123	0.05092661	0.0416138	0.04071903	0.03510794
Cumulative Proportion	0.57563682	0.62656343	0.6681772	0.70889627	0.74400421

	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15
Standard deviation	0.91848692	0.89907223	0.85466909	0.83749372	0.79365332
Proportion of Variance	0.03012922	0.02886896	0.02608783	0.02504985	0.02249591
Cumulative Proportion	0.77413343	0.80300239	0.82909022	0.85414007	0.87663598

	Comp.16	Comp.17	Comp.18	Comp.19	Comp.20
Standard deviation	0.70156640	0.69581758	0.64644816	0.63202274	0.60642489
Proportion of Variance	0.01757841	0.01729150	0.01492483	0.01426617	0.01313397
Cumulative Proportion	0.89421439	0.91150590	0.92643073	0.94069689	0.95383086

	Comp.21	Comp.22	Comp.23	Comp.24
Standard deviation	0.57312205	0.517253173	0.462898935	0.392913819
Proportion of Variance	0.01173103	0.009555387	0.007652694	0.005513617
Cumulative Proportion	0.96556190	0.975117283	0.982769977	0.988283594

	Comp.25	Comp.26	Comp.27	Comp.28
Standard deviation	0.351579330	0.305407180	0.264405901	0.203143615
Proportion of Variance	0.004414572	0.003331198	0.002496803	0.001473833
Cumulative Proportion	0.992698166	0.996029364	0.998526167	1.000000000

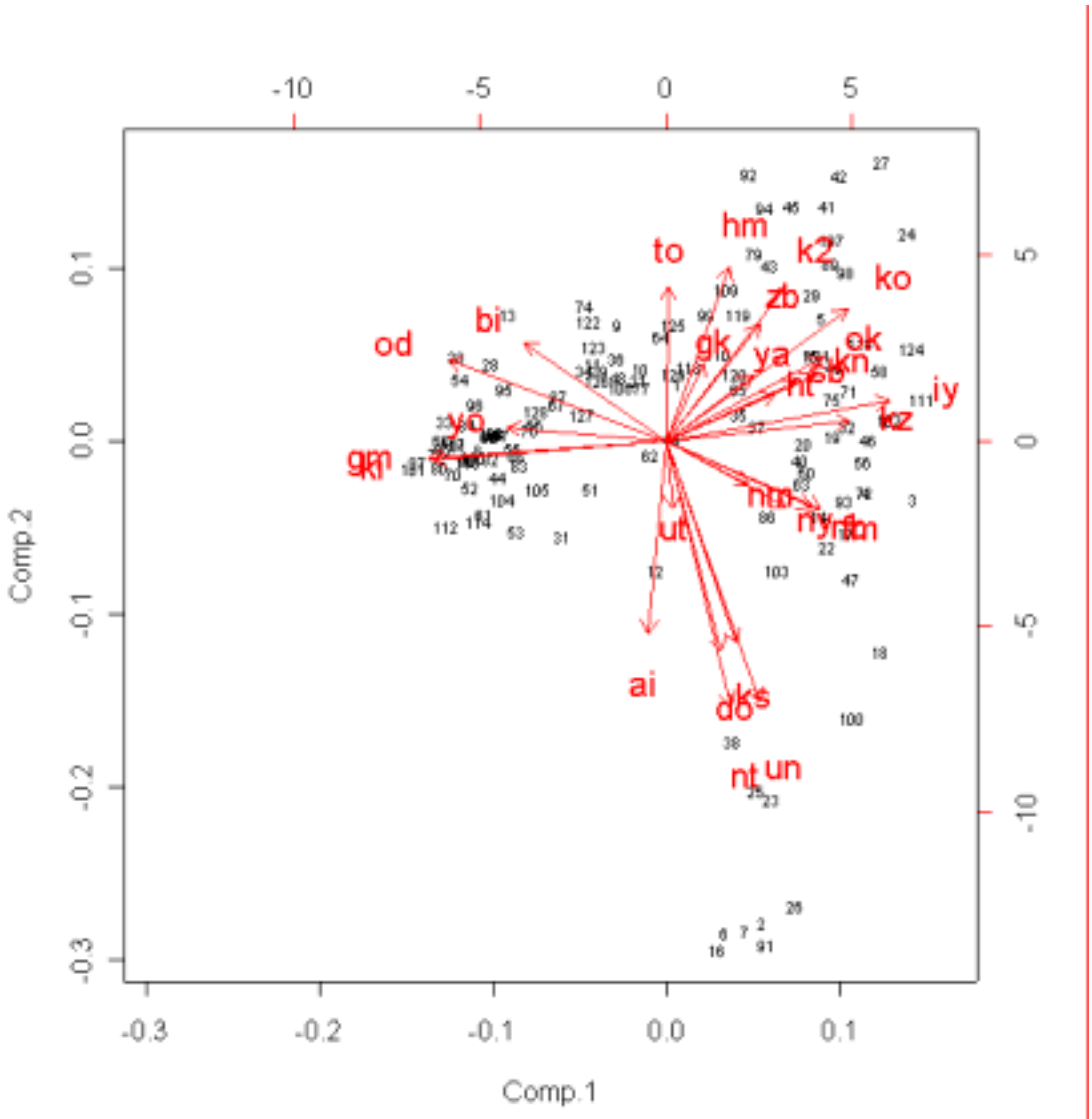

Figure 2. Bi-plot of the pca results for Japanese listeners. Numbers identify the individual speech tokens. Arrows show the strength and orientation of the categories in the first two principal component dimensions. Proximity of labels indicates closeness in the perceptual space. Length of arrows corresponds to a higher correlation or agreement in the responses. Inner products between variables approximate covariances and distances between observations approximate Mahalanobis distance.

Figure 2 plots the first two dimensions of the principal component analysis based on correlations between the labels produced by the Japanese listeners. The arrows show the strength and projected direction (in this two-dimensional view of the data space) of the categories determined by the labels. Individual speech tokens can be identified by their numeric identifiers positioned within the same view of the data space. The

length of the arrows corresponds to agreement in the number of responses for each label, so even though some areas of the space are only sparsely filled with points, a longer arrow in these areas indicates a stronger agreement amongst listeners on their choice of label for those utterances.

Table 8. Showing the first two loadings of each feature and explanations (with approximate translations) of the three-letter codes, to facilitate examination of the relative position of each component in the pca plot (figure 2). The initial "j" has been omitted in the plot for ease of viewing. Note that these loadings do NOT map directly to the numbers on the axes of the plot. They are included here for approximate guidance only.

-3.3	-0.3	jgm	gimon - question
-3.2	-0.3	jki	kikikaeshi - pardon?
-3.0	1.2	jod	odoroki - strong surprise
-2.2	0.2	jyo	yorokobi - pleasure
-2.0	1.5	jbi	bikkuri - surprise
-0.2	-2.9	jai	aizuchi - back-channel
0	2.3	jto	tomadoi - puzzled/bewildered
0.1	-1.0	jut	utagai - doubt
0.5	1.2	jgk	gakkari - disappointment
0.8	-3.2	jdo	doui - agreement
0.9	-4.0	jnt	nattoku - agreement
0.9	2.6	jhm	hirumu - flinch/shrink
1.0	-3.1	jks	kanshin - interest
1.2	1.0	jya	yabai - unpleasant
1.3	-4.0	jun	unazuki - nodding
1.3	1.8	jzb	zetsubou - despair
1.5	0.7	jht	hitei - disagree
1.7	-1.0	jny	nayamu - worried
1.7	2.3	jk2	kurushii - painful
1.8	0.9	jsb	shitsubou - depressed
1.9	1.2	jnm	(group:nemui mukanshin dousiyou) - (bored,uninterested,problem)
2.0	-1.0	jnb	nobasu - lengthened
2.1	1.0	jkn	kanashii - sad
2.2	1.3	jok	ochikomu - disappointed

2.2 -1.0 jfm fuman - frustration

2.5 2.0 jko (group:komaru,mayoi,tyuutyou) - (troubled,undecided,hesitation)

2.6 0.3 jkz kyozetsu - rejection

3.1 0.6 jiy iya - unpleasant

We can see from this figure that certain pairs of labels occupy very similar locations in the plot. For example, "gm" and "ki" (gimon = question, kikikaeshi = ask for repeat) are overlapping on the left of the plot, as are "do" and "ks" (doui = agreement, kanshin = interest) and "nt" and "un" (nattoku = understanding, unazaki = head nodding) at the bottom right. Since the listeners were free to choose their own terms to describe the categories, it is understandable that such pairs of synonyms should be found, and both encouraging and helpful for our analysis that the statistical processing should cluster them together on the basis of the observed correlations between the individual responses for each speech token, even though the responses appeared from the raw data to be far from unanimous.

Of further interest are the similarities calculated by the statistical model and indicated by the directions of the arrows in the plots. For instance, "kz" (kyouzetsu = rejection) and "iy" (iya = unpleasant) are close together but not overlapping. The similarity between these terms is clear, but one describes the act, and the other the attitude.

Similarly, we find that "ht" (hitei = disagreement) and "sb" (shitsubou = depressed) cluster below "kn" (kanashii = sad) and "ok" (ochikomu = disappointed), and that "nm" (nemui = sleepy) is somewhat more centralised but in the same quadrant as the cluster "nb" (nobashi = lengthening), "ny" (nayami = indecision), and "fm" (fuman = frustration). Clearly, listeners are responding to similar negative aspects in the speech, but they differ in terms of how they perceive and describe it. Some are more specific and others more general; some describe the speech, others the speaker, and yet others the speech act. However, we note that in many cases, the respondents appear to be broadly describing common or similar features. Not surprisingly, it appears to be rare for one listener to perceive a happy effect and another a sad one for the same speech utterance. We can therefore conclude that since the text of each utterance is the same (or would be transcribed using the same symbols) the manner of speaking is portraying the attributes of each utterance in a way that can be commonly understood by the majority of the listeners.

It is of interest here to speculate from the dispersion of the labels in the plot what these first two dimensions might represent. We see that the labels are well-distributed around the circle that describes the space of the distribution when plotted in these two dimensions, but note also that there are clear clusters and gaps. (in order to zoom in on certain parts of the figure to see the details of the clusters more clearly, it may be easier to access the electronic versions of these plots from our website.) The first component appears to map well onto "valency" with positive labels clustering towards the left of the figure, and negative ones to the right. The second dimension appears to represent “activation”; with labels clustered at the bottom of the figure representing inactive or passive states of the speaker, and those at the top indicating most active (the sign of the axes is not meaningful in principal component analysis, and should be ignored, since the axes merely represent abstract directions in a unitless coordinate space). While these interpretations must remain subjective and somewhat speculative, it is encouraging to find that the first two dimensions revealed by the objective statistical methods accord well with those proposed in the psychological literature [17].

Multi-cultural Perception

In order to test whether non-native foreign-language listeners were able to perceive the same or similar categories, we enlisted the help of listeners whose native language was not Japanese, and who professed no knowledge of that language. They were asked to label the same 129 instances of the utterance "eh" using the perceptual categories listed in Tables 3 and 4, which were suggested by the second author after a preliminary analysis of the results of the experiment with Japanese listeners. Two groups of listeners were employed; the first consisting of 12 people in Korea, and the second consisting of 19 in the United States of America.

Principal Component Analysis of these combined results provides interesting generalisations and shows that despite the negative impression caused by the confusion matrix display of the raw labelling scores, there is actually considerable agreement between the individual listeners and even between those of different linguistic backgrounds. When combining responses from Japanese (open-choice, 29 categories), Korean (closed choice, 17 categories), and American listeners (closed choice, 17 categories), we obtain a prediction model having 60 degrees of freedom. We find that the first principal component (or prediction coefficient) accounts for 31% of the variance in the data, and that the first five combined account for nearly 63% of the variance.

90% of the variance can be accounted for by just the first 20 principal components. There is therefore a considerable amount of overlap or redundancy between the labels, as would be expected if listeners of different language backgrounds hear the same or similar non-verbal characteristics in the speech sounds, since the English and Korean labels are expected to be a translation of the most commonly-used Japanese ones.

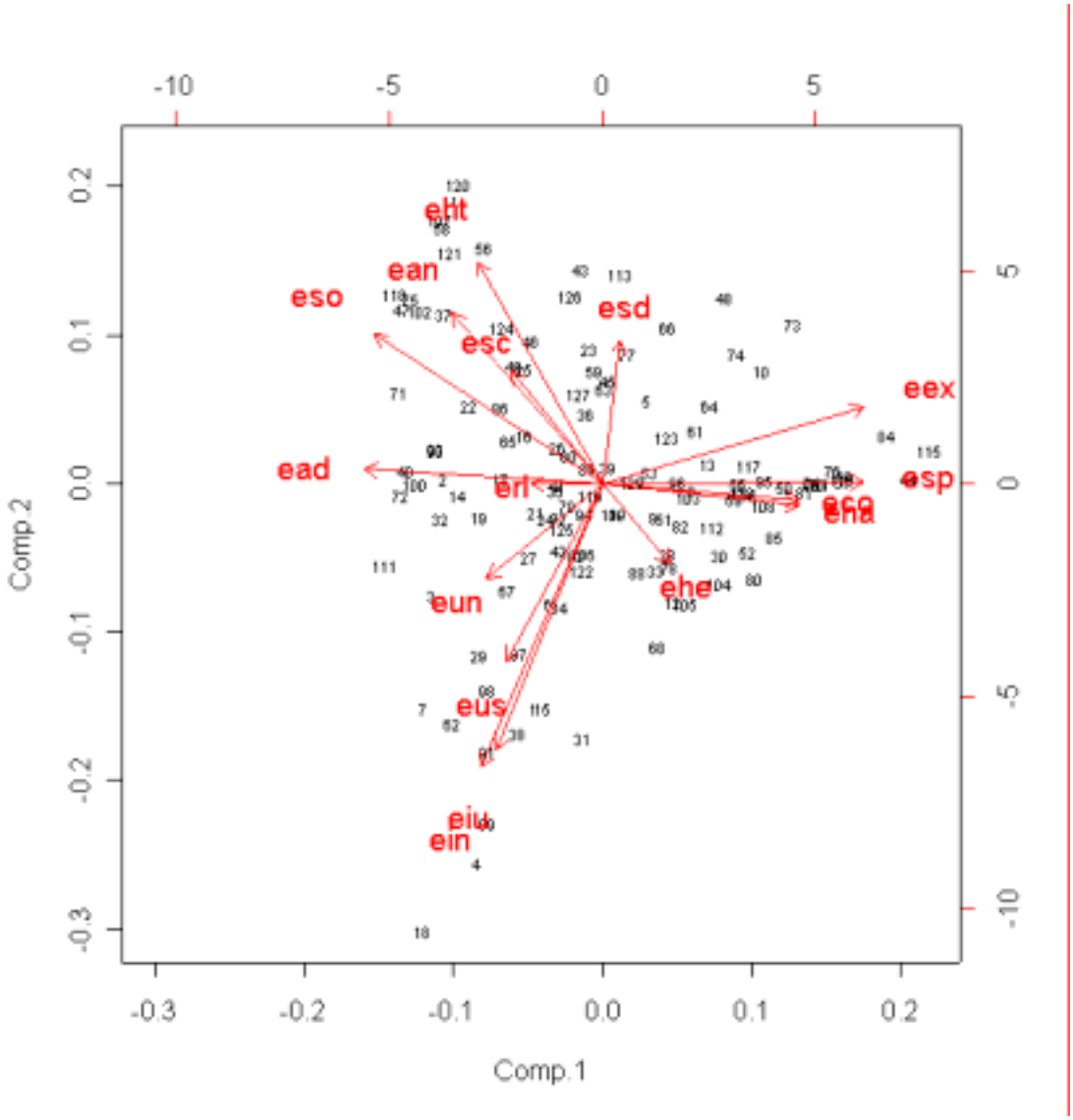

Figure 3. Biplot of the Principal Component Analysis results for American English listeners.. Numbers mark the speech tokens. Arrows show the strength and orientation of the features when mapped onto the first two principal component space. There is clear separation of the l;abels throughout the space, but we can see some clear overlaps where listeners have chosen different terms for essentially similar speech effects.

American English listeners

Figure 3 plots the results for the American English listeners. We can see that the axes appear to be rotated, with positive labels clustering towards the top-right of the plot, and negative ones towards the bottom-left. Labels marking strong effects are found at the top-left, and weak ones at the bottom right. This is directly mappable to the dimensions found for the analysis of Japanese responses.

Of interest here, though, is the overlap between *happy* ("eha") and *confused* ("eco") perceptions as shown on the right of the plot. At the bottom, we find a more expected overlap of *indifferent* ("ein") and *indifferent/unconcerned* ("eiu"), with *unsure* ("eus") and *unconcerned* ("eun") also appearing close nearby. The *hello* response ("ehe") is somewhat isolated in this projection of the data space, as is the *scared* ("esd") response. The cluster at the right includes *excited*, *surprised, happy* and *confused*, and that to the top-left *annoyed*, *sick-of*, *angry*, *scared*, and *hurt*, with *sad/crying* closer to the centre.

These responses show that there was a high degree of correlation between the choice of labels assigned to utterances having similar characteristics, and that whereas individual utterance-label combinations may be prioritised or ranked differently, the general trends appear to be the same. Listeners seem to generally hear the same effect from a given utterance. This in spite of the fact that none are familiar with the language or culture of Japanese. It may be that the paralinguistic signals in interjective speech are universal and cross language boundaries.

Korean listeners

Figure 4 shows the equivalent plot for results from the Korean listeners. We see that here too the first principal dimensions appear to represent valence and activation, this time (coincidentally) aligned similarly to those of the Japanese responses, with positive labels to the left, and negative ones to the right, and strong responses at the top, and weak ones at the bottom. However, the mapping is not identical.

We see that *questioning* clusters with *surprise* in the Japanese results, whereas *questioning* in Korean responses ("kgc") clusters on the other side of the plot with *annoyance* ("kch"). Whereas the *hello* response was isolated in the American English results, it clusters more closely to *other* ("kda") in the Korean results. However, it

should be noted that *hello* was a "late-added" item in the perception test, and was included as one of the items to be chosen for about half of the listeners of both languages. It would seem those Korean listeners who were not explicitly given the choice *hello* probably included these particular utterances in the *other*- category. American English listeners may not have felt so strongly about these utterances belonging to the *hello* category, and only when they were given the explicit choice, identified them as a category. As above, *happy*, *surprised*, and *excited* cluster together, as do *indifferent*, *unconcerned*, and *relieved*. However, whereas *unsure* ("kpl") appeared with *indifferent* in the American English responses, it clusters more closely with *hurt* in those of the Koreans, although the latter is more centralised in this case.

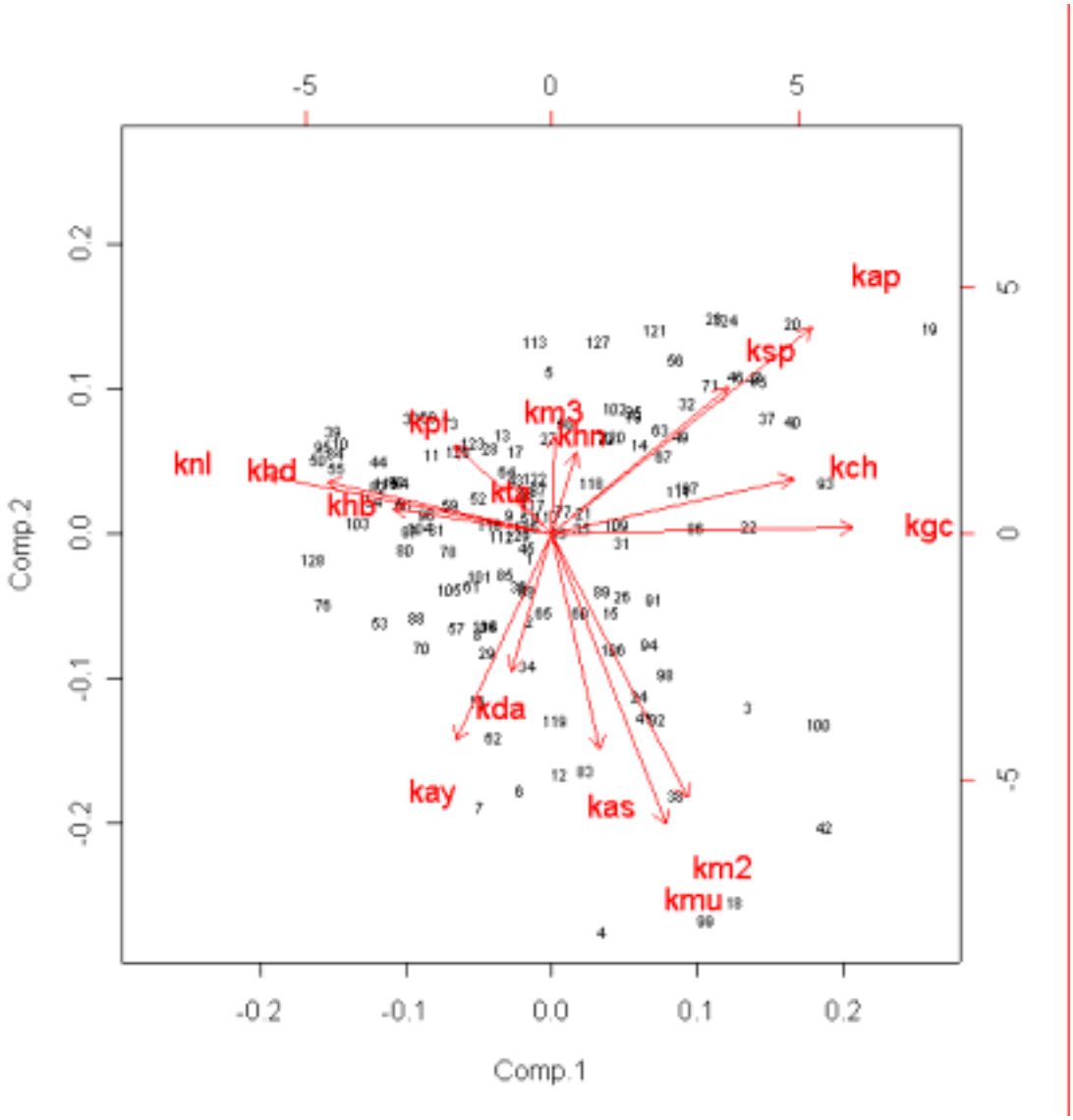

Figure 4. Biplot of the first two components from the Principal Component Analysis

results for Korean listeners. We can see similar overlaps to those of Figure 4.

Combined American English-Korean responses

Figure 5 plots the results of a pca analysis of the combined AE-Korean responses. We can identify five main groupings of responses, with *indifferenc*e at the top-left, *happy*, *surprised* and *excited* at the right, *hello* and *unsure* in between, *annoyed* and *confused* at the left, and *sick-of*, *angry,* and *scared* clustered together at the bottom.

The coincidence of locations for the *hello* response in this view of the data space is encouraging and confirms that the majority of both Korean and American English listeners heard these particular utterances more as a greeting than as anything else, whereas it is only in Korean that "Eh" is formalised linguistically (at least in the vernacular) as a greeting. No Japanese listener suggested this interpretation despite the free choice of lexical input. The category was inserted at the request of the Korean listeners in a preliminary trial, and included in the American English choices for compatibility.

Table 9 compares the percentage usage of each label with respect to the total number of labels assigned. It shows that whereas Korean listeners perceived 3.7% of utterances as type *hello*, only 0.9% of US responses were assigned to this category. However, whereas 5.5% of American English listeners perceived *indifference* ("ein"), only 1.2% of Korean responses were assigned to this category. The American English listeners did not perceive many utterances as showing *lack of concern* ("eun" = 1.7%) but the same category ("kmu") accounts for 5.2% of Korean responses. American English listeners perceived many more utterances to be *happy*. The majority Korean response was *apuda* ("kap" *sick-of*). We note that such differences might be cause for some international misunderstandings.

Table 9. Percentage of each label type with total number of labels assigned:

ead	ean	eco	eex	eha	ehe	eht	ein	eiu	erl	esc	esd	eso	esp	eun	eus
9.1	6.3	0.8	8.2	12.5	0.9	6.2	7.2	5.5	5.6	7.8	9.1	6.4	5.5	1.7	7.4
kgc	khn	kch	khd	khb	kay	kta	kmu	km2	kas	ksp	km3	kap	knl	kmu	kpl
8.7	5.9	1.8	5.5	6.2	3.7	7.5	5.2	1.2	8.5	6.8	6.2	9.6	7.9	5.2	5.7

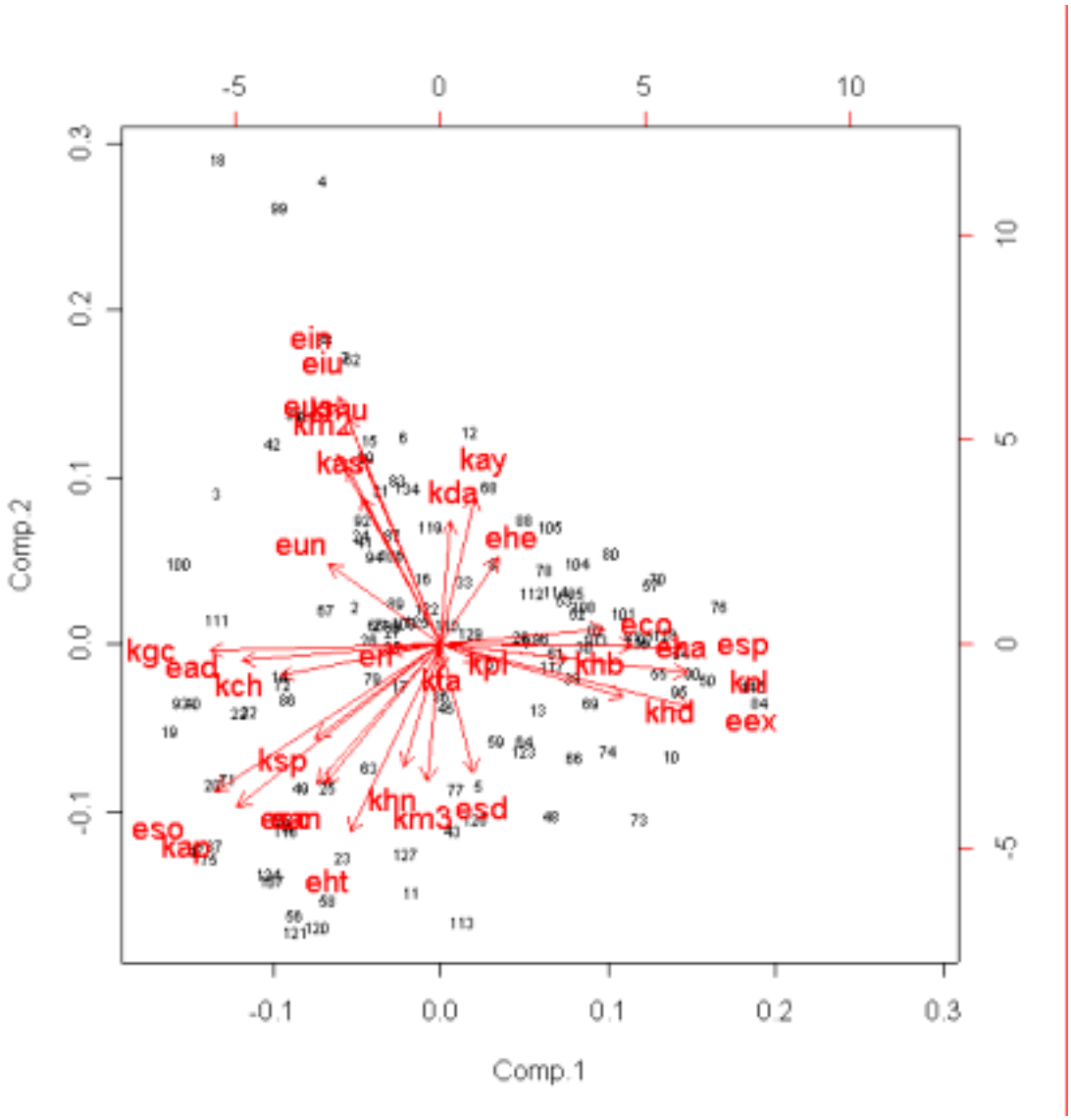

Figure 5. Biplot of the Principal Component Analysis results for both Korean and American English listeners combined. Here we can see that equivalent labels cluster together, confirming that in most cases, the Korean and American English respondents both hear essentially the same effects when listening to the different Japanese speech tokens.

Three listener groups combined

Finally, we plot the results for all three groups of listeners together. Figure 6 shows a bi-plot of results from a pca analysis of the combined responses for all listeners together. There is considerable overlap, but Table 10 provides a listing of the first two principal

component loadings (similar to coordinates in the 2-dimensioal space) for easier comparison. Labels having small differences in both dimensions of the loadings will appear together in the pca space and can be considered similar. We can see that there are many equivalences.

Table 10. Loadings for the first two Principal Components of the three-language analysis. Proximity in these coordinates indicates similarity of response from the listeners. We can see that equivalent or similar terms cluster more closely together.

Comp.1	Comp.2		+	Comp.1	Comp.2		+	Comp.1	Comp.2	
-22.8	-14.9	jiy	|	-9.1	7.7	km2	|	-0.3	-14.8	km3
-21.1	-8.7	eso	|	-9.0	4.0	eus	|	-0.3	-7.0	kta
-19.5	-12.0	kap	|	-8.8	-19.2	eht	|	-0.1	-10.5	jto
-19.3	0.9	kgc	|	-8.8	-4.8	jht	|	0.4	-12.5	esd
-18.0	9.4	ead	|	-8.2	4.8	jnm	|	3.4	0.9	kda
-17.7	10.1	jnb	|	-7.3	-7.6	jzb	|	3.5	16.5	kay
-17.6	-8.7	jkz	|	-7.3	24.9	kas	|	3.8	2.4	kpl
-17.2	-22.1	jko	|	-6.9	20.4	jks	|	5.4	9.7	ehe
-16.6	1.4	jfm	|	-6.7	27.8	jnt	|	10.1	4.1	khb
-15.4	8.4	jny	|	-6.7	-12.1	jya	|	12.8	-5.0	jbi
-13.8	1.8	kch	|	-6.6	20.8	eiu	|	14.1	-4.0	khd
-13.5	-3.7	jok	|	-6.5	15.1	kmu	|	15.2	3.1	eco
-13.1	-4.8	esc	|	-6.3	22.9	ein	|	17.1	-2.6	jyo
-12.6	-2.7	ksp	|	-5.9	27.0	jdo	|	18.7	-1.0	eha
-12.3	-6.5	jsb	|	-5.7	9.6	erl	|	20.2	-6.9	eex
-11.8	-4.6	jkn	|	-5.1	-12.0	jhm	|	20.9	-2.3	knl
-11.4	-17.0	ean	|	-2.9	-15.0	khn	|	21.7	7.3	jki
-11.1	32.6	jun	|	-1.9	-10.9	jgk	|	22.0	0.9	esp
-10.7	7.5	eun	|	-1.4	3.5	jut	|	22.1	6.4	jgm
-9.5	-12.5	jk2	|	-0.5	27.8	jai	|	22.8	-11.9	jod

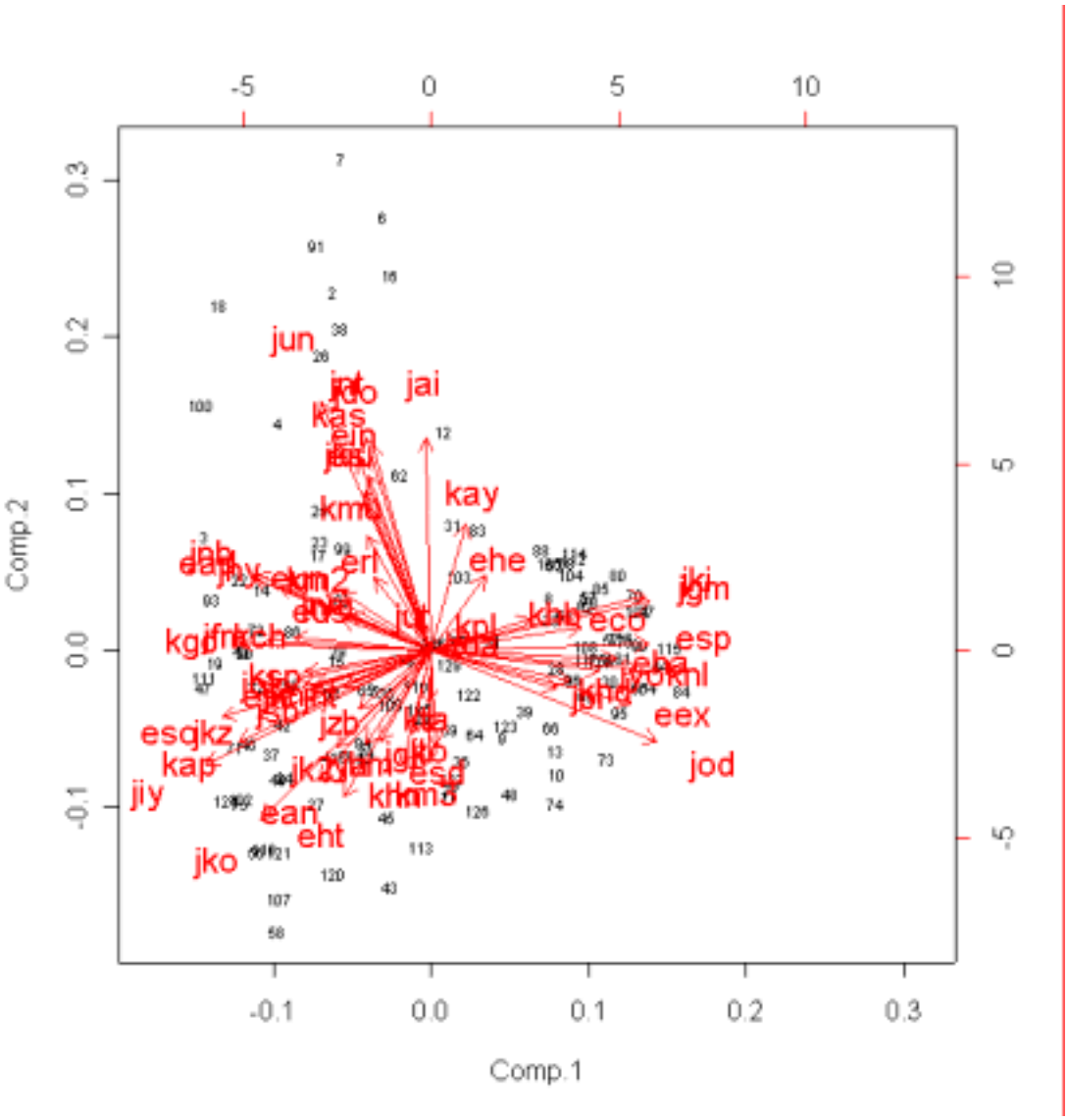

Figure 6. Bi-plot of the Principal Component Analysis results for three language-groups together. Labels starting with an "e" indicate responses from English-speaking listeners, those starting with a "k" from Korean listeners, and those starting with a "j" indicate responses from Japanese listeners. The latter are essentially the same as in Figure 1, but appear differently oriented as a result of the increased number of factors in the analysis. Table 10 provides the numerical coordinates for easier comparison.

Table 11 shows the amount of overall variance that is accounted for by the first fifteen components of the analysis. The first two components now account for almost half of the variance (44.1%, up from 29.9% for Japanese listeners alone).

Table 11. The contribution of the top 15 components in the multi-language pca analysis. We see that the first 10 components explain almost 80% of the variance, showing that there is a great amount of redundancy (i.e., a high degree of overlap) in the labels.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	5.0355702	3.2161866	2.33932915	2.26845195	2.13481363
Proportion of Variance	0.3132788	0.1277957	0.06761085	0.06357596	0.05630587
Cumulative Proportion	0.3132788	0.4410745	0.50868539	0.57226135	0.62856722
	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
Standard deviation	1.98790753	1.80704747	1.50137663	1.35966773	1.26016167
Proportion of Variance	0.04882318	0.04034343	0.02784922	0.02284017	0.01961943
Cumulative Proportion	0.67739040	0.71773384	0.74558306	0.76842323	0.78804265
	Comp.11	Comp.12	Comp.13	Comp.14	Comp.15
Standard deviation	1.19507370	1.09887065	1.06414946	1.02272812	1.01459316
Proportion of Variance	0.01764506	0.01491856	0.01399069	0.01292273	0.01271796
Cumulative Proportion	0.80568771	0.82060627	0.83459696	0.84751968	0.86023765

Moving counter-clockwise around the plot, we see first that "jai" (Japanese:*aizuch*i) appears in the most neutral position for component 1. The *aizuchi* is frequently used back-channel utterance in Japanese, which indicates to the speaker that the listener is paying attention, and signalling a mild request to continue speaking. It is perhaps the most unmarked of the “eh” variants in these data.

Next is a cluster "jdo", "jnt" (Japanese:*doui* and Japanese:*nattoku*) and "jks" (Japanese:*kanshin*), all signalling agreement, or a mild form of “yes”. These are interspersed with "ein" "eiu" and "kmu" (*indifferent* or *not carin*g in English and Korean), and to a lesser extent, as indicated by the shorter arrow, "erl" (*relieved*).

The next block, reading counter-clockwise in the plot (extreme in component 1, and slightly positive in component 2) includes "jnb" (nobasi = lengthening), "jny" (*nayami* = *undecided*), and "jnm" (*nemui* = *sleepy*), with "ead", "eun", "eus" (English:*annoyed*, *uncaring, unsure*), and "km2" (Korean:*indifferent/uncaring*).

The next block, on the extreme left of the plot (most negative in component 1, and

neutral in component 2) includes "jfm" (Japanese:*fuman* = *frustration*) "kgc" (Korean: quichanta = *annoyed/disgusted*), and (but closer to the centre of the plot, and therefore not so extreme) “kch” (Korean:*chilmunhada* = *questioning*)

Further down on the left of the plot, we find a cluster containing "jok" (*ochikomu kurai rakutan* = *disappointed*), "jht" (*hitei* = *disagreement*), "jsb" (*shitsubou* = *depressed*), "jkz" (*kyozetu* = *rejection*), "jkn" (*kanashii* = *sad*), "jiy" (*iya* = *unpleasant*), "esc" (*sad/crying*), "eso" (*sick-of*), "ksp" (*sad/crying*), and "kap" (*sick of)*. All are negative but relatively passive responses.

In the bottom left-hand corner of the plot, extreme in both components and with no opposing (top-right) counterpart, we find a group of "jko" (*komaru* = *to be troubled*), "jk2" (*kurushii* = *uncomfortable*), "jya" (*yabai* = *unpleasant*), "jzb" (*zetsubou* = *despair*), "ean" (*angry*), and "eht" (*hurt*). The similarity in meaning and placement of these labels is striking.

At the bottom of the figure, neutral in component 1 and extreme in component 2, we find "jto" (*tomadoi* = *bewildered*), "kta" (*hurt*), "khn" (*angry*), "km3" (*scared*), and "esd" (*sad*). The vertical dimension of the figure, (Component 2) appears to reflect “activation”, with weak or passive responses at the top, and stronger more expressive ones at the bottom.

Continuing on around the plot, there is then a gap until the relatively isolated set on the right of the figure, containing "jod" (strong *surprise*), "jbi" (*surprise*), "khd" (*excited*), "eex" (*excited*), followed by "jyo" (*yorokobi* = *pleasure*), "eha" (*happy*), "knl" (*surprised*), and then "kda" (other), "kpl" (*unsure*), "khb" (*happy*), "eco" (*confused*), "jgm" (*question*) , and "jki" (*clarification*). Finally, the pair of *hellos* sits alone in the top-right quadrant, both for Korean and American English listeners; no Japanese suggested this category.

Summarising, we tend to find clear positive responses (*happy, surprised, excited*) on the right of the plot, and more negative ones (*tired, angry, sad*) on the left. In the vertical dimension, we find weak or less personally involved responses (*indifference, agreement, relief*) at the top, and strong or involved ones (*excited, angry, sick-of*) at the bottom. The mapping between the first two principal component dimensions and the valency and activation of the psychological literature has been noted above for the Japanese data; it is encouraging to find it maintained even in the multicultural responses.

That agreement between listeners of different cultures and first languages is not absolute is not surprising--since speech interactions are by their very nature context-dependent. To ask listeners to label affective information from a single speech interjection is a very difficult task. The rather amazing results of this experiment indicate that it is possible for listeners to assign affective-type labels even to short, highly context-dependent speech utterances. Moreover, we see that clusters in the plot correspond to sense groups in the labels, regardless of language. From the principal component analyses, it appears that about 15 labels would be sufficient to describe this data.

We started this study with the assumption that the term "eh" is used in Japanese either as a backchannel, signifying mild agreement to indicate simply that the listener is paying attention to the speaker, or alternatively that it can be used in its strong form to mean lexical "yes". From our results, we now conclude that this term has a much broader function for providing affective information in discourse signals.

From the pca analyses, we conclude that the number of categories appropriate for the description of the functions of “eh” may be between fifteen and twenty. It is clear that the majority of listeners perceive *indifference, relief, surprise, annoyance, confusion, anger, concern,* etc., but it is also clear that they are not unanimous in their selection of the best descriptor. For the American English responses, we note from Table 3 that the counts of a descriptor along the main diagonal were sometimes different than expected, i.e., for *confused* more listeners preferred excited or happy; for *hello*, these were also labeled as e*xcited, happy,* or *indifferent*; *surprise* as *excited* or *happy*; and *uncaring* as *happy* or *indifferent*. The Korean responses were stronger along the main diagonal (no competing label was preferred), but the responses were far from unanimous. We believe that this is not an indicator of ambiguity in the pronunciations so much as an indicator that people prioritise their perceptions differently, some responding more to the speech act, others to the speaker-state, yet others to the discourse intent. The topic of individual variation in prioritising percepts is one that requires further research. In addition, different listeners most likely respond differently to different aspects of the acoustic signal—some listeners may be more sensitive to changes in pitch, while others loudness, and still others, voice quality. We hope to pursue this line of investigation as part of our ongoing study of perception of affective information in conversational speech. It is likely that there is no “one-right-answer” for any individual utterance, but that a

vector of responses, each having different activations, would better describe the perception process. Listeners probably perceive all aspects of the affective information and it is the nature of the (artificial) labelling task that requires them to focus only on the top one or two. The software interface that we used for this experiment has proven useful in that it has shown us what terms can be used to describe the set of utterances, and has provided data that, taken together, shows us the set of descriptors that apply to each utterance. Speech technology applications will be able to utilise a feature-vector of activations, but in the present paper, it would make for some very difficult reading since, by nature, it cannot be simplified.

Acknowlegments

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. We wish to thank Ms.Jean Hee Jung for her invaluable help in soliciting Korean listeners. We also thank the students at Black Hills State University for their help with the perception tests.

References

[1] Maekawa, K. (1998). Phonetic and phonological characteristics of paralinguistic information in spoken Japanese, Proc. ICSLP98 (CD-ROM), Paper #0997

[2] Mozziconacci, S., (2002) “Prosody and emotions”, in Proc Speech Prosody.

[3] Ekman, P., (1992) “An argument for basic emotions”, 169-200 in Stein et al (eds) Basic Emotions, Lawrence Erlbaum.

[4] Campbell & Mokhtarri, (2003) “Voice Quality, the 4th prosodic dimension”, in Proc ICPhS, Barcelona.

[5] ESP pages: http://www..feast.his.atr.jp

[6] Hayashi, Y. (1998) 「音声に含まれる感性的情報とピッチ曲線 感動詞『ええ』を利用して」ー日本音響学会講演論文集 pp.381-382.

[7] Hayashi, Y. (1999). Recognition of vocal expression of emotion in Japanese: Using the interjection ‘eh’. International Congress of Phonetic Sciences, San Francisco, pp. 2355-2358.

[8] Campbell, (2003) ”Voice characteristics of spontaneous speech”, Proc ASJ Spring meeting.

[9] Erickson, D. and Maekawa, K. (2001). Perception of American English emotion by Japanese listeners. *Acoustical Society of Japan, Spring Meeting*.

[10] Erickson, D., Ohashi, S., Makita, Y., Kajimoto, N., Mokhtari, P. (2003). Perception of naturally-spoken expressive speech by American English and Japanese listeners. Proceedings of The 1st JST/CREST International Workshop on Expressive Speech Processing, Kobe, Feb. 21,22, 2003, pp. 31-36.

[11] Campbell, N., (2002) "Recording techniques for capturing natural everyday speech", in Proc LREC.

[12] Campbell, N., (2004) "The JST Expressive Speech Corpus", in Proc LREC.

[13] Campbell, N., (2004) "Listening between the lines; a study of paralinguistic information carried by tone-of-voice" in Proc TAL, Beijing, China.

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

[15] Mardia, K. V., J. T. Kent and J. M. Bibby (1979). Multivariate Analysis, London: Academic Press.

[16] Venables, W. N. and B. D. Ripley (1997, 9). Modern Applied Statistics with S-PLUS, Springer-Verlag.

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

Okay – here's where it stops … for now …