



# Morphology of Vocal Affect Bursts: Exploring Expressive Interjections in Japanese Conversation

Hiroki Mori

Graduate School of Engineering, Utsunomiya University, Japan

hiroki@speech-lab.org

## Abstract

Expressive interjection (EI) is defined as non-lexical speech sound which indicates the speaker’s cognitive/affective state changes. It is a type of vocal affect burst, i.e., brief and sudden nonverbal expressions that are produced spontaneously and unconsciously. Although EI as a social signal is assumed to play an important part in speech communication, very little is known about its linguistic, paralinguistic, and pragmatic nature. The goal of this study is to unveil the structure and functions of vocal affect bursts in human interactions. This paper focuses on the surface structure of EIs, which is an indispensable foundation for further analysis and modeling. Based on linguistic/acoustic analyses of a natural, spontaneous dialog corpus, the distinctiveness of EI was revealed as: (1) less variation in transcribed expression, (2) may have very short duration, and (3) higher F0 and intensity. In addition, it was revealed that an apparent correlation between formant frequencies and perceived paralinguistic information was observed only for EIs with the vowel /a/, which suggests that the vowel /a/ as an EI can accommodate richer paralinguistic information than other vowels.

**Index Terms:** affect bursts, interjections, nonverbal, paralinguistic information, dimension

## 1. Introduction

*Affect burst* [1, 2] refers to brief and sudden nonverbal emotional expressions in face and voice. Laughter is a typical example of an affect burst. Although the definition of affect burst varies from literature to literature, a common notion may be that affect bursts are produced more or less spontaneously and unconsciously. Figure 1 shows the vocal phenomena that are considered in this paper to be affect bursts. In the definition here, not only *raw* affect bursts [1] such as laughing, screaming and crying voice, which are totally not under the control of the speaker, but also *affect emblems* including conventionalized symbols such as “kya:” (Japanese onomatopoeia of a scream) and affective lexical items, which are somewhat under control, are regarded as affect bursts. Despite their wide spectrum, every affect burst can be characterized as a social signal that forms another communication channel and provides information regarding the speaker’s mental or cognitive states.

This paper features so-called *expressive interjection* (EI) [3], a type of affect burst. It is defined in this paper as non-lexical speech sounds which indicate the speaker’s cognitive/affective state changes. EI is somewhat prelinguistic, but it can be transcribed. Therefore, EI can be positioned halfway between raw affect burst and affect emblem (Fig. 1). Although EI is assumed to play an important part in speech communication, very little is known about its linguistic, paralinguistic, and pragmatic nature.

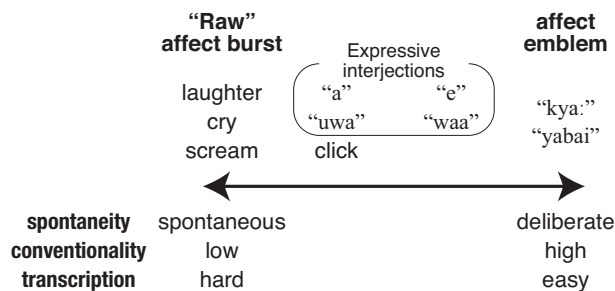


Figure 1: Spectrum of affect bursts.

This paper explores the linguistic and paralinguistic structure of EI in Japanese conversation. Based on analyses of a natural, spontaneous dialog corpus, the following prediction is verified: *Expressive interjection is a distinctive entity, not only in its form but also in its function as a carrier of the speaker’s state.*

## 2. Corpus

The Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UU Database) [4] is a collection of natural, spontaneous dialogs of college students. This corpus is especially intended for use in understanding the usage, structure and effect of paralinguistic information in expressive Japanese conversational speech. The task of the dialogs, namely “four-frame cartoon sorting,” was carefully designed to stimulate expressively-rich and vivid conversation. In this task, four cards each containing one frame extracted from a cartoon are shuffled, and each participant has two cards out of the four, and is asked to estimate the original order without looking at the remaining cards. The task proved to motivate the participants quite well because Japanese students like cartoons and were eager to know the true story.

The current release of the UU Database includes dialogs of seven pairs of college students (12 females, 2 males). The participants and pairings were selected carefully to ensure that both people in each pair were of the same grade and able to get along well with each other. All of the participants used “tameguchi (lit. equal speech)” style, an informal speaking style commonly used in everyday conversation between close friends or relatives, in speaking to the partner. In total, dialog speech was recorded for 27 sessions, lasting about 130 minutes. The whole speech signal was segmented into 4840 utterances, then further segmented into chunks (stretches of speech sound), short pauses, and non-speech sounds such as laughter.

10.21437/Interspeech.2015-326

In the UU Database, each utterance is assigned a six-dimension vector, as shown below, that describes different aspects of paralinguistic information perceived from the utterance.

- (1) pleasant-unpleasant
- (2) aroused-sleepy
- (3) dominant-submissive
- (4) credible-doubtful
- (5) interested-indifferent
- (6) positive-negative

The first three dimensions can be regarded as equivalent to the Valence-Arousal-Dominance [5] (or Activation-Evaluation-Power [6]), a de facto standard for dimension-based emotion description in recent studies on speech and emotion.

After a screening test, three qualified annotators evaluated perceived paralinguistic information for each utterance on a 7-point scale. In evaluating the pleasant-unpleasant scale, for example, 1 corresponds to extremely unpleasant, 4 to neutral, and 7 to extremely pleasant.

In order to analyze linguistic and acoustic features, chunks that were judged as EIs were tagged by the author, in addition to existing tags. Their timings were also determined manually. The basic criteria used to identify EIs were as follows:

- An EI is a non-lexical speech sound which indicates the speaker’s cognitive/affective state changes.
- EIs are *not* optional, unlike fillers.

### 3. Are Expressive Interjections Different from Fillers?

Filler [7] is another type of interjection. It is similar to EI in some aspects, e.g. it signals the speaker’s mental process, some filler is non-lexical, etc. The Corpus of Spontaneous Speech (CSJ) [10] does not distinguish EI from filler, and assigns the tag (F) to both of them. However, typical filler cannot be regarded as an affect burst, and it is very likely that EIs can be distinguished from fillers morphologically or pragmatically.

In this section, linguistic and acoustic patterns of EIs and fillers are analyzed. All of the following analyses are for female pairs in the UU Database, comprising 23 sessions.

#### 3.1. Usage

Figure 2 shows the frequency distribution of EIs, in comparison to fillers and other chunks. Every speaker uses EIs to some extent. This implies that, as expected, affect burst is an indispensable part of speech communication, just like disfluencies such as fillers [8]. The usage of filler is, however, more speaker-dependent than EIs. Some speakers (e.g. FTS) use a lot; others (e.g. FMS) use almost none.

Figure 3 shows the number of appearances of EIs and fillers. Transcribed expressions of EI have less variability compared to that of filler. The top-three expressions “a” “e” and “N” alone account for more than 80 percent of all EIs.

Morphologically, it is an interesting issue whether an EI and a filler that have an identical transcription can be distinguished acoustically. For example, “e” [9] is an expression that can be interpreted both as an EI and as a filler. Although it is the most frequent filler in monologue speech in CSJ, “e” as a filler appeared only eight times in the UU Database, partly because the database contains casual conversations rather than academic presentations in CSJ. Likewise, there is no major EI that shares

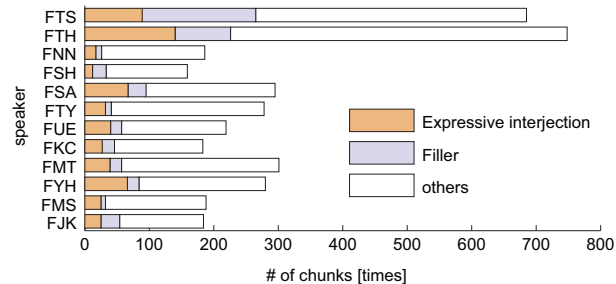


Figure 2: Frequency distribution of EIs, fillers, and other chunks in the UU Database.

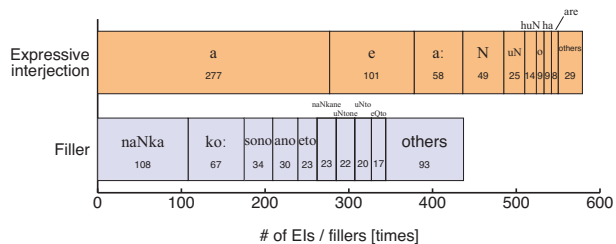


Figure 3: Frequency distribution of transcription for expressive interjections and fillers in the UU Database.

the same expression with a filler. Because of this, the comparison described in the following subsection is performed without considering the transcription.

#### 3.2. Acoustic Features

##### 3.2.1. Duration

Figure 4 shows the distribution of duration for EIs and fillers. The duration of EIs is distributed over quite a wide range, similar to that of fillers. Some of them are almost as long as 2000 ms. The mean duration is almost the same (EI: 338 ms, filler: 337 ms), which is much longer than that of single vowels (76 ms). But, the shape of their distribution is significantly different (Mann-Whitney *U* test,  $U = 101205$ ,  $p < 0.01$ ). The median is 205 ms for EIs and 291 ms for fillers. This suggests that EIs can be produced not only as a long vocalization but also, unlike fillers, as a very brief vocalization. The prosodic variety of EI might reflect its functional difference (i.e. filler-like vs. burst-like), which needs to be investigated in future.

##### 3.2.2. F0 and intensity

As prosodic features, mean F0 and peak intensity were calculated for monosyllabic EIs and fillers. F0 values were obtained with the *To Pitch (ac)* command of Praat, and converted to the mel scale. To avoid errors due to unstable phonations, the mean calculation was performed only for ones with five voiced frames or more. Likewise, intensity values were obtained with the *To intensity* command of Praat.

Figure 5 shows (a) mean F0 and (b) peak intensity distribution for EIs and fillers. This indicates that EIs were produced with higher pitch and louder voice as a whole. The mean F0 of EIs was significantly higher than that of fillers (two-sample *t*-test,  $t(536) = 4.07$ ,  $p < 0.01$ ), and the peak intensity of EIs was significantly higher than that of fillers (Welch’s *t*-test,  $t(388.9) = 4.35$ ,  $p < 0.01$ ).

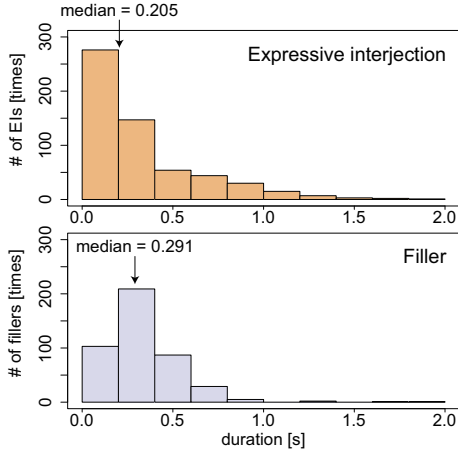


Figure 4: Duration distribution of expressive interjections and fillers in the UU Database.

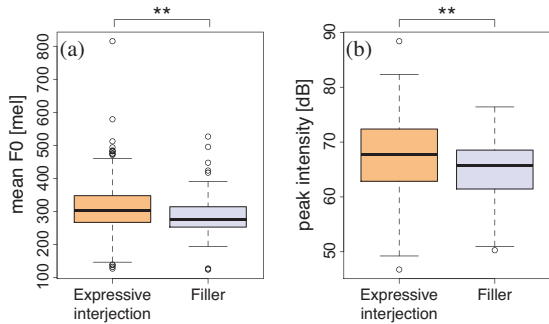


Figure 5: Parameter distribution of expressive interjections and fillers in the UU Database. (a) mean F0, (b) peak intensity.

However, there were a considerable number of EIs that were produced with a low pitch and quiet voice. This is not surprising, because affect burst can be observed when a speaker is experiencing a low-arousal emotion such as disappointment or anxiety. The broader distribution of acoustic features coincides with the prediction that EI can be a carrier of various paralinguistic information.

### 3.2.3. Spectral tilt

Vowel spectral tilt was calculated as a voice quality feature related to the glottal source. The first LPC cepstrum coefficient ( $c_1$ ) was obtained with the *To LPC (autocorrelation)* command of Praat for the midpoints (50% of the duration) of EIs and fillers, then log power at 0 Hz relative to 3000 Hz was calculated using the spectral envelope estimated by  $c_1$ .

Spectral tilt differences between EI and filler were highly dependent on speakers. In the following analysis, speakers that did not produce fillers more than three times were excluded. A two-way ANOVA (type [EI or filler] (2)  $\times$  speaker (9)) revealed a significant main effect of type ( $F(1, 453) = 8.44, p < 0.01$ ) and speaker ( $F(8, 453) = 10.18, p < 0.01$ ). Also, a significant interaction between type and speaker was found ( $F(8, 453) = 6.63, p < 0.01$ ). Figure 6 shows the spectral tilt distribution for each speaker. For three speakers out of nine, significant simple

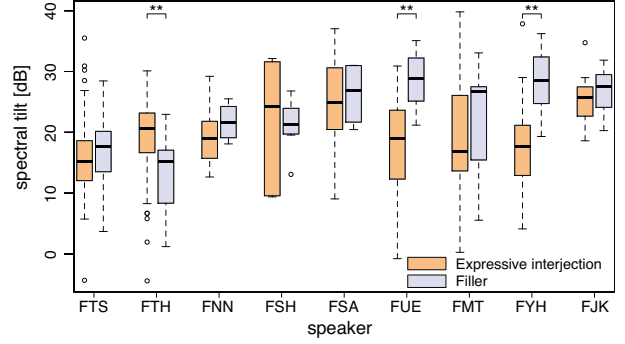


Figure 6: Per-speaker distribution of spectral tilt for expressive interjections and fillers in the UU Database.

Table 1: Mean ratings of emotion dimensions for the expressive interjections “a” “e” and “N.”

pleasantness	a (4.47) > e (3.81) = N (3.60)
arousal	a (5.12) = e (5.05) > N (4.05)
dominance	e (4.65) > a (4.37) = N (4.12)
credibility	a (4.90) > N (3.73) = e (3.71)
interest	a (5.38) = e (5.33) > N (4.72)
positivity	a (4.98) > N (3.78) = e (3.74)

main effects of type were found ( $p < 0.01$ ), though they were not consistent. For speakers FUE and FYH, spectral tilt of EIs was shallower; in other words, EIs were produced with greater vocal effort than fillers. On the contrary, speaker FTH produced EIs with steeper spectral tilt.

## 4. Expressive Interjection as a Carrier of Paralinguistic Information

### 4.1. Expression and Emotion Dimensions

As described in Section 2, utterances in the UU Database are given dimension-based ratings that describe emotional and attitudinal information that are perceived. Although they are meant to describe paralinguistic information rather than linguistic content, it is possible that the ratings more or less reflect speakers’ choice of words. Therefore, the choice of EI expression might have some relationship to the emotion dimensions.

Here, the overall ratings of the three most typical EIs “a” “e” and “N” are compared. Table 1 shows the mean ratings of the three EIs for the six dimensions. With one-way ANOVA, a significant effect of EI expression was found for all the dimensions ( $p < 0.01$ ). Table 1 also shows the result of post-hoc tests (Tukey’s pairwise comparison), where > indicates a significant difference between the pair ( $p < 0.05$ ). From this table, a clear contrast among these EIs can be observed: (1) “a” is perceived as much more pleasant, more credible, and more positive than others, (2) “N” is perceived as less aroused and less interested than others, and (3) “e” is perceived as more dominant than others. These results represent quite well the emotional/attitudinal effect of EIs intrinsic to expressions.

### 4.2. Paralinguistic Variation of Acoustic Features

Our previous study [4] investigated acoustic correlates (F0, intensity, etc.) of perceived paralinguistic information. Figure 7 shows the relationship between paralinguistic information and acoustic features, focusing on EIs. From this figure, it can be understood that perceived paralinguistic information of EIs is

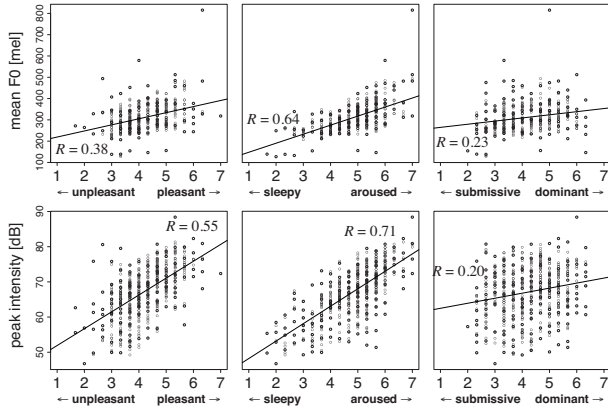


Figure 7: Relationship between paralinguistic information and acoustic features (upper: mean F0, lower: peak intensity).

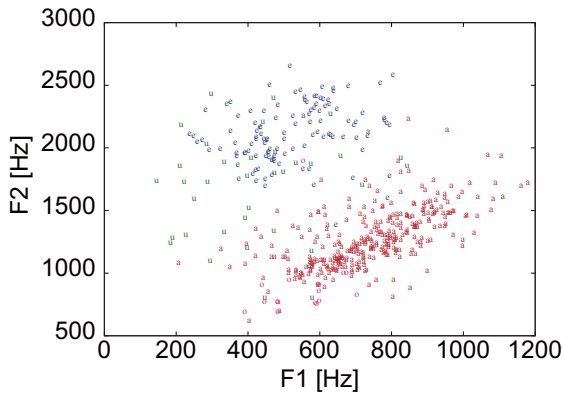


Figure 8: Formant frequency distribution of expressive interjections.

distributed over a wide range, implying that EIs are a powerful carrier of paralinguistic information. It is also observed that the mean F0 and peak intensity of EIs account for a considerable portion of the variation of pleasantness and arousal (but not dominance).

Segmental features can also reflect emotion-related information associated with EI. First and second formant frequency ( $F_1$ ,  $F_2$ ) were calculated for the same set analyzed in Section 3 with the *To Formant (burg)* command of Praat for the midpoints of EIs, then erroneous values were discarded by manual inspection. Figure 8 shows the distribution of  $F_1$ - $F_2$  plotted by the letters indicating nuclear vowel. The formant frequency is distributed over a wide range, even for the same vowel. This is partly due to the speaker variation, but it may also be a paralinguistic effect of EIs.

To prove this, the relationship between paralinguistic information and formant frequency was inspected. Figure 9 shows the distribution of  $F_1$  and  $F_2$  with respect to two fundamental emotion dimensions: pleasant-unpleasant and aroused-sleepy. Regression lines indicate that the regression coefficient for the vowel is significant ( $p < 0.05$ ). Apparently, the correlation is highly dependent on vowels. For EIs with the vowel /a/, both  $F_1$  and  $F_2$  well reflect the paralinguistic information, namely, EIs with higher formant frequency are perceived as more pleas-

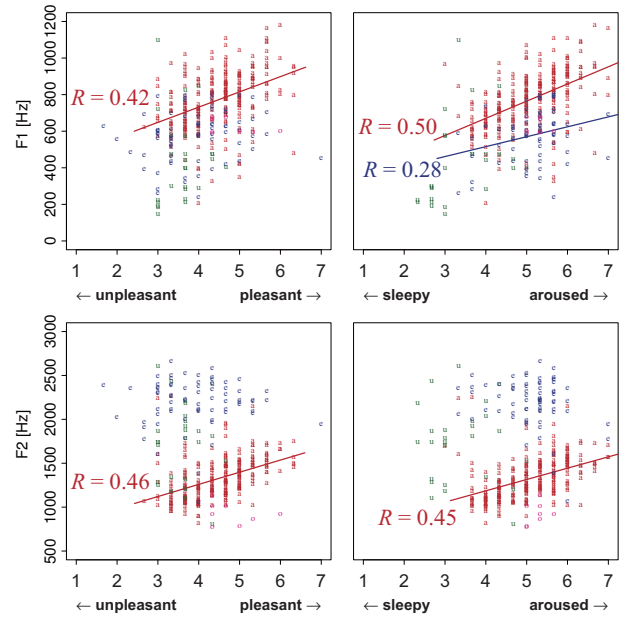


Figure 9: Distribution of  $F_1$  (upper) and  $F_2$  (lower) of expressive interjections with respect to paralinguistic information.

ant and more aroused. This can be explained as follows. High  $F_1$  reflects vowel openness [11], which is consistent with vocalization in high arousal. In addition, high pleasantness is often associated with a smiling face, which causes higher formant frequencies due to retraction of the lip corners [12].

On the other hand, correlation was not significant for other vowels, except the one between  $F_1$  and arousal for /e/. This suggests that EIs with /a/ are very special, as they can accommodate richer paralinguistic information than ones with other vowels.

## 5. Conclusions

In this paper, the linguistic and paralinguistic structure of expressive interjections in Japanese conversation was explored. EIs may have very short duration, unlike fillers. In addition, F0 and intensity of EIs are higher than fillers as a whole. These results support the assumption that EIs are distinguished from filler. Compared with fillers, EIs have less variability in transcribed expression. Nevertheless, the paralinguistic variation of EIs is large enough to convey rich paralinguistic information, especially with the vowel /a/.

These findings provide a foundation for future corpus development or advanced speech technologies. For example, to develop a dialog system that can produce affect bursts using corpus-based speech synthesis technology, EIs in the corpus should be annotated with a dedicated tag.

The next step is to explore the effect of EIs in speech communication, and to model the relationship between structure and effect of EIs, which will hopefully lead to a better understanding of nonverbal aspects of speech communication.

## 6. Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 26280100, 26284062.

## 7. References

- [1] K. R. Scherer, "Affect bursts," in *Emotions: Essays on emotion theory*. Hillsdale, New Jersey: Lawrence Erlbaum, 1994, pp. 161–193.
- [2] M. Schröder, "Experimental study of affect bursts," *Speech Communication*, vol. 40, no. 1-2, pp. 99–116, 2003.
- [3] F. Ameka, "Interjections: The universal yet neglected part of speech," *Journal of Pragmatics*, vol. 18, pp. 101–118, 1992.
- [4] H. Mori, T. Satake, M. Nakamura, and H. Kasuya, "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Commun.*, vol. 53, pp. 36–50, 2011.
- [5] D. Wu, T. Parsons, E. Mower, and S. S. Narayanan, "Speech emotion estimation in 3D space," in *Proc. ICME*, 2010, pp. 737–742.
- [6] M. Schröder, R. Cowie, and E. Douglas-Cowie, "Acoustic correlates of emotion dimensions in view of speech synthesis," in *Proc. Eurospeech 2001*, 2001, pp. 87–90.
- [7] H. H. Clark, "Managing problems in speaking," *Speech Commun.*, vol. 15, no. 3-4, pp. 243–250, Dec. 1994.
- [8] M. Watanabe, Y. Den, K. Hirose, S. Miwa, and N. Minematsu, "Factors affecting speakers' choice of fillers in Japanese presentations," in *Proc. Interspeech 2006*, 2006, pp. 1256–1259.
- [9] N. Campbell and D. Erickson, "What do people hear? A study of the perception of non-verbal affective information in conversational speech," *J. Phonet. Soc. Jpn.*, vol. 8, no. 1, pp. 9–28, 2004.
- [10] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proc. ISCA & IEEE Wkshp. Spontaneous Speech Processing and Recognition (SSPR2003)*, 2003, pp. 7–12.
- [11] R. D. Kent and C. Read, *The Acoustic Analysis of Speech*, 2nd ed. San Diego, CA: Singular, 2001.
- [12] V. C. Tartter and D. Braun, "Hearing smiles and frowns in normal and whisper registers," *Journal of the Acoustical Society of America*, vol. 96, no. 4, pp. 2101–2107, 1994.