Accuracy of Automatic Cross-Corpus Emotion Labeling for Conversational Speech Corpus Commonization

Hiroki Mori[†], **Atsushi Nagaoka**[†], **Yoshiko Arimoto**[‡§]
[†]Graduate School of Engineering, Utsunomiya University, Japan
[‡]Brain Science Institute, RIKEN, Japan
[§]College of Engineering, Shibaura Institute of Technology, Japan
{hiroki,atusi}@speech-lab.org, ar@brain.riken.jp

Abstract

There exists a major incompatibility in emotion labeling framework among emotional speech corpora, that is, category-based and dimension-based. Commonizing these requires inter-corpus emotion labeling according to both frameworks, but doing this by human annotators is too costly for most cases. This paper examines the possibility of automatic cross-corpus emotion labeling. In order to evaluate the effectiveness of the automatic labeling, a comprehensive emotion annotation for two conversational corpora, UUDB and OGVC, was performed. With a state-of-the-art machine learning technique, dimensional and categorical emotion estimation models were trained and tested against the two corpora. For the emotion dimension estimation, the automatic cross-corpus emotion labeling for the different corpus was effective for the dimensions of aroused-sleepy, dominant-submissive and interested-indifferent, showing only slight performance degradation against the result for the same corpus. On the other hand, the performance for the emotion category estimation was not sufficient.

Keywords: emotion, spontaneous speech, emotion recognition, openSMILE, support vector machine

1. Introduction

Large-scale spoken language resources are essential for understanding and modeling speech and other human behaviors. A wide variety of corpus-based speech technologies have been developed for recognizing, reproducing or manipulating verbal and nonverbal behaviors in human communication. Among them, affective computing is one of the most prominent research trends. Research goals of affective computing include analysis and modeling of speech emotion, recognition or detection of the speaker's affective state (Schuller et al., 2013), and emotional speech synthesis (Nagata et al., 2013).

Although several speech corpora have been developed for studying speech emotion, it is widely recognized that less acted, more realistic data are needed (Schuller et al., 2013). However, developing a naturalistic emotional corpus tends to suffer from the problem of scalability, mainly due to the costly work of annotating emotion.

If multiple emotional corpora could be put together into a large corpus, it would benefit many studies and applications. However, the framework used by each corpus for describing emotion is generally incompatible with each other. A major incompatibility comes from the theory of emotion assumed, i.e. category-based and dimension-based (Cowie and Cornelius, 2003). The former assumes some emotion categories such as the "Big Six" emotions (fear, anger, happiness, sadness, surprise, disgust), while the latter does not assume discrete categories but rather regards an emotional state as a point in a space of a small number of dimensions. Emotional speech corpora with category-based labels include Berlin Database of Emotional Speech (Emo-DB) (Burkhardt et al., 2005), Parameterized & Annotated CMU Let's Go (LEGO) Database (Schmitt et al., 2012), Surrey Audio-Visual Expressed Emotion (SAVEE) Database (Haq and Jackson, 2010), FAU Aibo Emotion Corpus (Steidl, 2009) and Online Gaming Voice Chat Corpus with Emotional Label (OGVC) (Arimoto et al., 2012), while those with dimension-based labels include Vera am Mittag (VAM) German Audio-Visual Spontaneous Speech Database (Grimm et al., 2008) and Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies (UUDB) (Mori et al., 2011). The Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database (Busso et al., 2008) contains both categorical and dimensional labels.

This paper examines the possibility of automatic cross-corpus emotion labeling. Suppose we have two conversational speech corpora with different annotation schema. If the emotion labels of corpus A can be reliably estimated based on corpus B's annotation framework, we will virtually obtain a large corpus composed of A+B, with a common annotation framework. In principle, this can be achieved by training a model of emotional speech from the utterances of corpus B, employing some machine learning method.

For a prototype of the cross-corpus emotion labeling problem, two specific Japanese conversational corpora are used in the present study, i.e. UUDB as a dimension-based corpus, and OGVC as a category-based corpus. Figure 1 illustrates the concepts of automatic cross-corpus emotion labeling. A regression model that predicts dimensional description from speech parameters is trained from UUDB, which thereafter is applied to OGVC to estimate the dimensional description of emotion (b). Likewise, a classification model that predicts categorical description is trained from OGVC, then is applied to UUDB to estimate the categorical description of emotion (d). Our interest here is how accurately the model can estimate emotions of different corpora (b & d) compared to the same corpus (a & c), using a standard but state-of-the-art feature extraction and machine

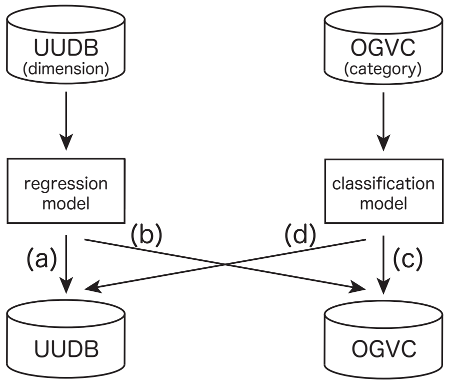


Figure 1: An illustrative example of cross-corpus emotion labeling.

learning method for emotion recognition. This paper does not consider ways to improve the method itself.

With the present corpora, the accuracy of (b) and (d) cannot be evaluated because of the lack of "correct labels" given by human annotators: UUDB does not have categorical annotations, and OGVC does not have dimensional annotations. To make the evaluation possible, new annotators were recruited for this study to evaluate emotions for both UUDB and OGVC, with both annotation frameworks.

2. Corpora

2.1. UUDB

UUDB (Utsunomiya University, 2008) is a collection of natural, spontaneous dialogs of Japanese college students. The participants were engaged in the "four-frame cartoon sorting" task, where 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 current release of the UUDB includes dialogs of seven pairs of college students (12 females, 2 males), composed of 4840 utterances.

For all utterances, perceived emotional states of speakers are provided. The emotional states were annotated with the following six abstract dimensions:

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

After a screening test, three qualified annotators evaluated the emotional state 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.

2.2. OGVC

OGVC (Arimoto and Kawatsu, 2012) is a speech corpus containing spontaneous dialogue speech and its emotional labels. The naturalistic emotional speech part contains 9114 spontaneous utterances from five dyadic and one triad dialogues. In the recording, a massively multi-player online role-playing game (MMORPG) was used to stimulate players to express emotion. The participants were 13 college students (4 females, 9 males) with experience of playing online games.

After screening out speakers with low sound levels and utterances that could not be transcribed, the total number of utterances was 6578. The emotional state of each utterance was annotated by three labelers. They had to choose one emotional state to label each utterance from ten alternatives of eight emotional states: fear (FEA), surprise (SUR), sadness (SAD), disgust (DIS), anger (ANG), anticipation (ANT), joy (JOY), and acceptance (ACC), as well as a neutral state (NEU) and others (OTH). The eight emotional states were selected with reference to the primary emotions in Plutchik's multidimensional model (Plutchik, 1980).

3. Methods

3.1. Inter-corpus Emotion Annotation

To obtain reliable and consistent emotion annotation across labelers, a screening test was conducted with 54 utterances each from UUDB and OGVC. Ten labelers participated in the screening test, and three out of the ten were selected.

The main work consists of the annotation for 4840 utterances of UUDB and 6578 utterances of OGVC. The annotation framework was basically the same as described in Sections 2.1. (dimension) and 2.2. (category). The selected three evaluators conducted both dimensional annotation and categorical annotation for all 11418 utterances from both corpora. In the following analyses, the newly obtained emotion annotation described in this section is exclusively used; the original dimensional annotation of UUDB and categorical annotation of OGVC will not be used.

The evaluation of the three annotators had to be unified in order to use machine learning. For the dimensional evaluation, we simply averaged the evaluated values. Figure 2 shows the distribution of the averaged values for the dimensions of pleasant-unpleasant and aroused-sleepy. The distribution was similar for the two corpora, with a slight difference in the upper-left region: OGVC contains more utterances with unpleasant and aroused emotional states (such as anger or fear) than UUDB.

For the categorical evaluation, the unification is less straightforward. We decided to extract an "agreed" subset from the whole dataset: If the category labels of the three annotators for an utterance were all different, the utterance was marked as "unagreed" and not included in the subset. Afterwards, the major category was regarded as the "correct label" for the utterance. Table 1 shows the number of emotion categories that are assigned to the utterances of the "agreed" subset. It is understood that emotion categories are unevenly distributed both in OGVC and UUDB.

3.2. Automatic labeling

Acoustic features of utterances were extracted using openSMILE (Eyben et al., 2010), with the preset configuration for the Interspeech 2010 Paralinguistic Challenge baseline system (Schuller et al., 2010). The number of dimensions

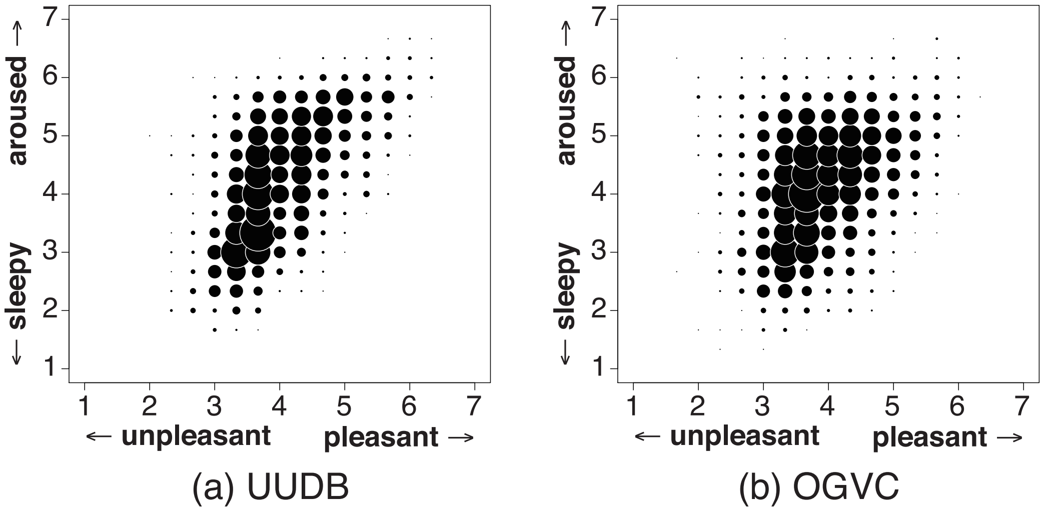


Figure 2: Distribution of dimensional evaluations for the dimensions of pleasant-unpleasant and aroused-sleepy. The circle area is proportional to the number of occurrences.

Table 1: Numbers of the major (two or more votes) emotion categories (“agreed” subset).

	ACC	FEA	SUR	SAD	DIS	ANG	ANT	JOY	NEU	total
OGVC	623	282	311	488	969	128	186	438	18	3443
UUDB	1030	94	120	331	406	39	59	259	13	2351

was 1582. No speaker adaptation or channel adaptation technique was employed in this study.

The regression models for estimating emotion dimensions, based on the Support Vector Regression (Bishop, 2006), were trained with UUDB. The classification model for estimating emotion categories, based on the Support Vector Classification, were trained with OGVC. Both types of models were built with the kernlab package of R. In evaluating the regression model against UUDB (a) and the classification model against OGVC (c), the leave-one-speaker-out cross-validation was performed, whereas in evaluating the regression model against OGVC (b) and the classification model against UUDB (d), simple open tests were performed. For the dimensional estimation, the accuracy was evaluated with the Pearson correlation coefficient R and the root mean square error (RMSE). For the categorical estimation, the accuracy was evaluated with the Weighted Average Recall (WAR) and Unweighted Average Recall (UAR). The WAR is the correct classification rate for the entire test set:

$$\mathrm{WAR} = \frac{|\{i \in I | H_i = R_i\}|}{|I|}, \quad (1)$$

where I is the test set, R_i is the correct label for the utterance i, and H_i is the estimated label for the utterance i, whereas the UAR is the correct classification rate averaged over the emotion categories:

$$\mathrm{UAR} = \frac{1}{|K|} \sum_{k \in K} \frac{|\{i \in I | R_i = k \wedge H_i = R_i\}|}{|\{i \in I | R_i = k\}|}, \quad (2)$$

where K is the set of emotion categories.

Table 2: Accuracy of emotion dimension estimation for (a) same corpus, and for (b) different corpus.

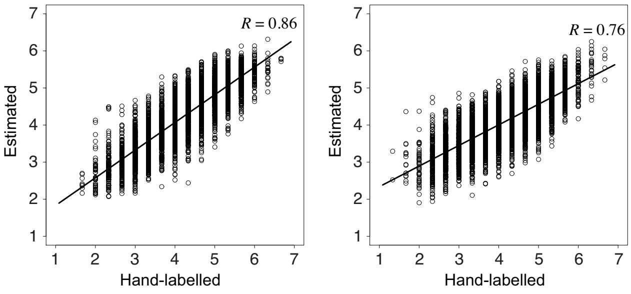
	(a)		(b)	
	R	RMSE	R	RMSE
pleasant-unpleasant	0.64	0.51	0.34	0.63
aroused-sleepy	0.86	0.51	0.76	0.56
dominant-submissive	0.85	0.57	0.70	0.62
credible-doubtful	0.63	0.53	0.35	0.62
interested-indifferent	0.78	0.46	0.61	0.54
positive-negative	0.43	0.54	0.19	0.61

4. Results and Discussion

4.1. Emotion Dimension Estimation

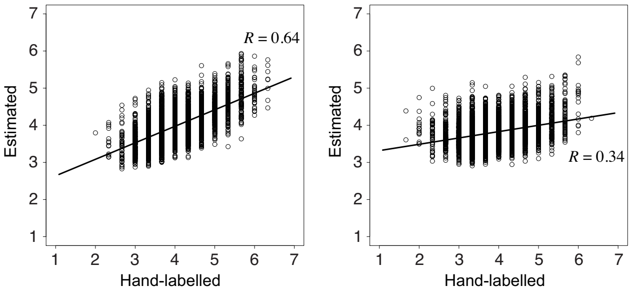
Table 2 shows the accuracy of emotion dimension estimation for UUDB ((a) in Fig. 1) and for OGVC ((b) in Fig. 1). As can be seen from (a) in Table 2, quite a high correlation was obtained for the same corpus, especially for the dimensions of aroused-sleepy, dominant-submissive, and interested-indifferent. The accuracy for these dimensions was also high even for the different corpus, as shown in (b) in Table 2. Figure 3 shows the scatter plots of hand-labeled and estimated arousal for (a) the same corpus and for (b) the different corpus. The estimation was reasonably accurate both for the same corpus and for the different corpus.

On the other hand, the correlation for the different corpus was much lower than that for the same corpus for the dimension of positive-negative. Similarly, the accuracy for the different corpus was not sufficient for the dimensions of pleasant-unpleasant and credible-doubtful. Figure 4 shows



(a) UUDB (b) OGVC

Figure 3: Scatter plots of hand-labeled and estimated emotion (the dimension of aroused-sleepy).



(a) UUDB (b) OGVC

Figure 4: Scatter plots of hand-labeled and estimated emotion (the dimension of pleasant-unpleasant).

the scatter plots of hand-labeled and estimated pleasantness. Compared to the result for the same corpus (a), the model did not properly estimate pleasantness for the different corpus (b), especially for unpleasant utterances. One possible cause of this is the different distribution of the two corpora, as UUDB contains fewer utterances with pleasantness around 2 (very unpleasant) than OGVC, as shown in Fig. 2.

These results lead to the conclusion that automatic cross-corpus labeling of emotion dimensions may be effective for some, but not all, dimensions.

4.2. Emotion Category Estimation

Table 3 shows the accuracy of emotion category estimation for OGVC ((c) in Fig. 1) and for UUDB ((d) in Fig. 1). For reference, the baseline accuracy (majority voting) was 28.1% WAR and 11.1% UAR for the same corpus (c), and 43.8% WAR and 11.1% UAR for the different corpus (d). Although the accuracy was higher than the baseline, the performance of emotion classification was not sufficient even for the same corpus (c). The discrepancy between WAR and UAR was caused at least partly by the imbalance of emotion category in OGVC (see Table 1). An additional attempt to reweight instances for equalizing the importance of emotion categories (Yang et al., 2005) resulted in a slight (< 5 %) improvement in UAR at the expense of a slight degradation in WAR.

Table 3: Accuracy of emotion category estimation for (c) same corpus, and for (d) different corpus.

(c)		(d)	
WAR [%]	UAR [%]	WAR [%]	UAR [%]
30.3	21.7	34.1	22.6

At the first glance of Table 3, the performance for the different corpus (d) was higher than that for the same corpus, but in fact this merely reflected the imbalance of emotion category (Table 1). To investigate the classification results in detail, the confusion matrices are shown in Fig. 5. The confusion patterns for the same corpus (c) and different corpus (d) were similar. For example, the accuracy for sadness, surprise, disgust and acceptance was relatively high, while many utterances were misrecognized as disgust, acceptance and joy. This tendency seems to reflect the relatively large number of training samples of these categories, as shown in the upper row in Table 1.

5. Conclusions

In this paper, the effectiveness of automatic cross-corpus emotion labeling was evaluated by performing a comprehensive inter-corpus emotion annotation. With a state-of-the-art machine learning technique, dimensional and categorical emotion estimation models were trained and tested against two conversational corpora, UUDB and OGVC. For the emotion dimension estimation, the automatic cross-corpus emotion labeling was effective for some dimensions, showing only slight performance degradation. On the other hand, we could not obtain sufficient performance for the emotion category estimation.

We emphasize that cross-corpus emotion labeling is meant to help commonizing speech corpora, so it can be semi-automatic rather than fully automatic. Manual labeling would be required only for utterances whose emotion labels estimated by machine learning were not reliable. Therefore, predicting confidence for estimated emotion labels is an important issue to be addressed in future.

6. Acknowledgements

This work was supported by Tateishi Science and Technology Foundation Grant for Research (A), and JSPS KAKENHI Grant Number 26280100.

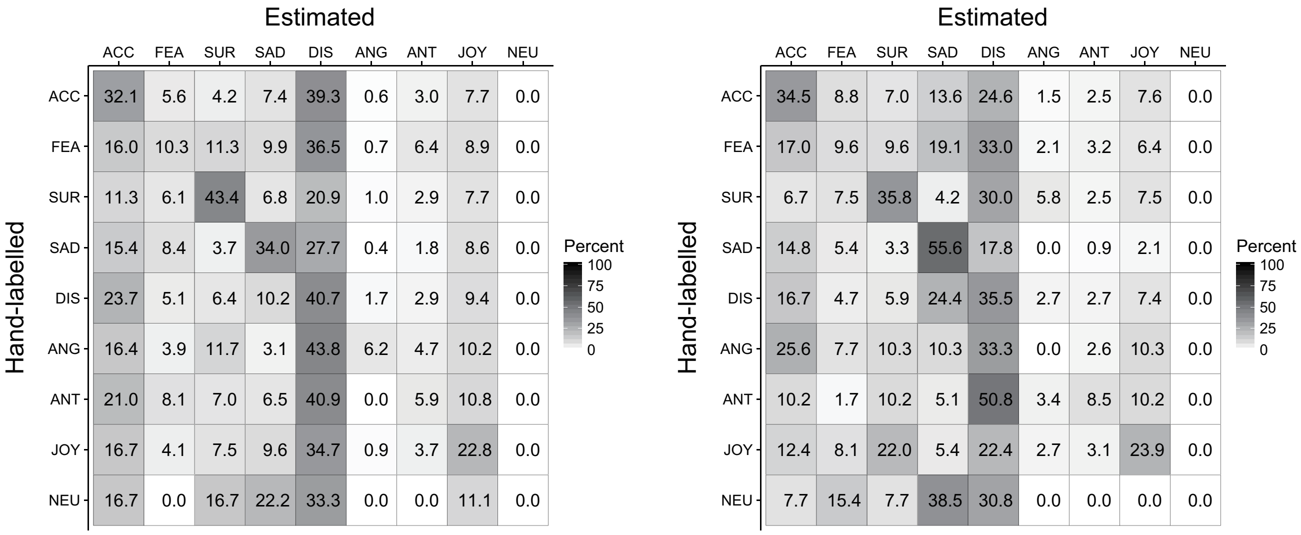
7. Bibliographical References

Arimoto, Y., Kawatsu, H., Ohno, S., and Iida, H. (2012). Naturalistic emotional speech collection paradigm with online game and its psychological and acoustical assessment. *Acoustical Science and Technology*, 33(6):359–369.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag, Secaucus, NJ, USA.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of German emotional speech. In *Proc. of Interspeech 2005*, pages 3–6.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.



Hand-labelled \ Estimated	ACC	FEA	SUR	SAD	DIS	ANG	ANT	JOY	NEU
ACC	32.1	5.6	4.2	7.4	39.3	0.6	3.0	7.7	0.0
FEA	16.0	10.3	11.3	9.9	36.5	0.7	6.4	8.9	0.0
SUR	11.3	6.1	43.4	6.8	20.9	1.0	2.9	7.7	0.0
SAD	15.4	8.4	3.7	34.0	27.7	0.4	1.8	8.6	0.0
DIS	23.7	5.1	6.4	10.2	40.7	1.7	2.9	9.4	0.0
ANG	16.4	3.9	11.7	3.1	43.8	6.2	4.7	10.2	0.0
ANT	21.0	8.1	7.0	6.5	40.9	0.0	5.9	10.8	0.0
JOY	16.7	4.1	7.5	9.6	34.7	0.9	3.7	22.8	0.0
NEU	16.7	0.0	16.7	22.2	33.3	0.0	0.0	11.1	0.0

(c) OGVC

Hand-labelled \ Estimated	ACC	FEA	SUR	SAD	DIS	ANG	ANT	JOY	NEU
ACC	34.5	8.8	7.0	13.6	24.6	1.5	2.5	7.6	0.0
FEA	17.0	9.6	9.6	19.1	33.0	2.1	3.2	6.4	0.0
SUR	6.7	7.5	35.8	4.2	30.0	5.8	2.5	7.5	0.0
SAD	14.8	5.4	3.3	55.6	17.8	0.0	0.9	2.1	0.0
DIS	16.7	4.7	5.9	24.4	35.5	2.7	2.7	7.4	0.0
ANG	25.6	7.7	10.3	10.3	33.3	0.0	2.6	10.3	0.0
ANT	10.2	1.7	10.2	5.1	50.8	3.4	8.5	10.2	0.0
JOY	12.4	8.1	22.0	5.4	22.4	2.7	3.1	23.9	0.0
NEU	7.7	15.4	7.7	38.5	30.8	0.0	0.0	0.0	0.0

(d) UUDB

Figure 5: Confusion matrices for emotion category estimation (in percent).

Cowie, R. and Cornelius, R. R. (2003). Describing the emotional states that are expressed in speech. *Speech Communication*, 40(1–2):5–32.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). openSMILE — The Munich versatile and fast open-source audio feature extractor. In *Proc. ACM Multimedia*, pages 1459–1462.

Grimm, M., Kroschel, K., and Narayanan, S. (2008). The Vera am Mittag German audio-visual emotional speech database. In *Proc. ICME 2008*, pages 865–868.

Haq, S. and Jackson, P., (2010). *Machine Audition: Principles, Algorithms and Systems*, chapter Multimodal Emotion Recognition, pages 398–423. IGI Global, Hershey PA.

Mori, H., Satake, T., Nakamura, M., and Kasuya, H. (2011). Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics. *Speech Communication*, 53(1):36–50.

Nagata, T., Mori, H., and Nose, T. (2013). Robust estimation of multiple-regression HMM parameters for dimension-based expressive dialogue speech synthesis. In *Proc. Interspeech 2013*, pages 1549–1553.

Plutchik, R. (1980). *Emotions: A Psychoevolutionary Synthesis*. Harper & Row, New York.

Schmitt, A., Ultes, S., and Minker, W. (2012). A parameterized and annotated spoken dialog corpus of the CMU Let's Go bus information system. In *Proc. LREC 2012*, pages 3369–3373.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Mueller, C., and Narayanan, S. (2010). The Interspeech 2010 Paralinguistic Challenge. In *Proc. Interspeech 2010*.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S. S. (2013). Paralinguistics in speech and language: State-of-the-art and the challenge. *Computer Speech and Language*, 27(1):4–39.

Steidl, S. (2009). *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag, Berlin.

Yang, X., Song, Q., and Cao, A. (2005). Weighted support vector machine for data classification. In *Proc. IJCNN '05*, volume 2, pages 859–864.

8. Language Resource References

Arimoto, Y. and Kawatsu, H. (2012). *Online gaming voice chat corpus with emotional label*. Distributed via NII-SRC.

Utsunomiya University. (2008). *Utsunomiya University Spoken Dialogue Database for Paralinguistic Information Studies*. Distributed via NII-SRC, Release 1.