

EMOGIB: Emotional Gibberish Speech Database for Affective Human-Robot Interaction

Selma Yilmazyildiz, David Henderickx, Bram Vanderborght,
Werner Verhelst, Eric Soetens, and Dirk Lefebber

Interdisciplinary Institute for Broadband Technology (IBBT), Belgium,
Dept. of Electronics and Informatics (ETRO - DSSP), Vrije Universiteit Brussel,
Belgium

Dept. of Cognitive Psychology, Vrije Universiteit Brussel, Belgium
Dept. of Mechanical Engineering (MECH - RMM), Vrije Universiteit Brussel, Belgium
{syilmazy,david.henderickx,bram.vanderborght,
wverhels,eric.soetens,dlefeber}@vub.ac.be
<http://www.ibbt.be/>,<http://www.vub.ac.be/>

Abstract. Gibberish speech consists of vocalizations of meaningless strings of speech sounds. It is sometimes used by performing artists or by cartoon animations (e.g.: Teletubbies) to express intended emotions, without pronouncing any actually understandable word. The facts that no understandable text has to be pronounced and that only affect is conveyed create the advantage of gibberish in affective computing. In our study, we intend to experiment the communication between a robot and hospitalized children using affective gibberish. In this study, a new emotional database consisting of 4 distinct corpuses has been recorded for the purpose of affective child-robot interaction. The database comprises speech recordings of one actress simulating a neutral state and the big six emotions: anger, disgust, fear, happiness, sadness and surprise. The database has been evaluated through a perceptual test for all subsets of the database by adults and one subset of the database with children, achieving recognition scores up to 81%.

Keywords: emotional speech database, emotional speech corpus, affective speech, gibberish speech, human-computer interaction

1 Introduction

Everyone would have heard a small baby communicating with his or her mother. But most likely not many would have paid attention to how smoothly they communicate their emotions without saying any single meaningful word. Although it is common knowledge that small children are able to do this, this effect is hard to replicate technically.

Like in the communication of babies with their mothers, a nonsense language like gibberish can be a successful carrier to express emotions and affect. Moreover, since there is no meaningful content and the focus of the listener is entirely on the conveyed affect, gibberish might even be more effective than meaningful

speech. This is the main motivation to use affective gibberish speech for communication between robots and children in our study.

In our previous study [4], the experiments concluded that gibberish speech can convey the emotions as effectively as semantically neutral speech. This supports our intention to use gibberish speech to express the emotions of the robots. In that study [4], to produce gibberish speech, we developed a program that replaces the vowel nuclei in a text with other vowel nuclei of the same language such that the text loses its meaning. We then used the generated gibberish text as input for TTS engines to produce the gibberish speech. But there are two drawbacks of this method. The final expressive speech strongly depends on the TTS engine quality and the voice quality of the emotions in the database is lost. To overcome these drawbacks, we decided to use a data-driven method that starts with a gibberish emotional database.

The lack of databases with genuine interaction is a key challenge in the studies of emotion expression. Observational or post hoc analyses of human interaction data is a method that could be used but it is a fairly impractical route to choose. As a result in most of the currently available databases acting has been used [1]. Busso and Narayanan argue that the methodologies and materials used to record the existing corpora are the main problem with the existing databases and not the use of actors itself. Some of the important requirements that need to be carefully considered in the design of the database are the speaker selection, contextualization and social setting, utilization of acting styles, usage of trained actors and the definition of the emotional descriptors [2].

However, the usage of affective gibberish speech and targeting the primary usage of the database in communication between robots (such as Probo[12] and NAO) and children help to simplify some of these requirements. First of all, there is no context in the gibberish speech. Secondly, Moris theory of the uncanny valley suggests that when robots look and act almost like actual humans, it causes a response of revulsion among human observers. The "valley" in the uncanny valley hypothesis represents a dip in the positivity of human reaction as a function of a robot's lifelikeness [3]. We can deduce from the uncanny valley theory that when the children notice a certain level of acting or unnaturalness in the synthesized speech of the robot will not necessarily negatively affect their overall communication experience with these robots.

2 EMOGIB-Emotional Gibberish Speech Database

EMOGIB is an expressive gibberish speech database that contains approximately 15 minutes of speech (\sim 1800 words) for each big six emotions (anger, disgust, fear, happiness, sadness, surprise) and 25 minutes of speech (\sim 4100 words) for neutral state. It has 4 different gibberish corpuses: C1 & C3 - generated by using the whole consonant and vowel space of Dutch and English, C2 & C4 - generated by using the whole vowel space and voiceless consonant space of Dutch and English. The reason of generating C2 & C4 comes from the ease of using voiceless consonants for automatic segmentation and manipulation.

2.1 Speaker Selection

Many of the requirements that effects the quality of the final database are influenced by the acting qualities of the selected speaker. Even though it is possible to improve the performance of the speaker by carefully designing the recording conditions[2], the speaker selection is still a key factor.

A call for speakers was distributed to the theater/drama schools in the country. Six of the candidates were invited for a phone interview. The candidates were all informed before the interview that they would be asked to voice-act in the interview. We sent them four sentences (one in English, one in Dutch and two nonsense sentences) that might be used as scripts to voice-act.

The interview started with a friendly talk where we asked their personal information such as their name, age, study program, languages spoken, experience in voice acting, experience in communication with children. The questions in the second part were structured in a way that we could evaluate the candidates on the following criteria: the ability to easily switch the voice to another type, the ability to act emotions, the ability to act nonsense sentences, the flexibility of the voice, the duration of the recording session, the capability of maintaining the voice quality during the recording session and the ability to act as fitting the required characteristics. We described them certain characteristics of an imaginary robot (such as *humor, pleasure, funny, stupid, emotional, sympathetic*) and asked them to speak spontaneously as if being one of those robots. This was to evaluate their ability to easily switch the voice to another type and their ability to act as fitting the required characteristics. To judge their ability to act emotions and their ability to act nonsense sentences, we instructed them to act the scripts that we sent them in six basic emotions (*happiness, sadness, fear, surprise, anger, and disgust*). Finally, to assess the flexibility/limits of their voice, we requested them to act in certain ages and genders such as *male, female, child, old man, old lady*. All the interview sessions were conducted through an Alcatel-Lucent 4019 phone in hands-free mode and the sessions were recorded to be able to listen to them later for evaluation.

Based on the above criteria, a 20 year old female drama student was selected as the speaker for the actual recordings.

2.2 Text Corpus

Languages consist of ruled combinations of words and words consist of specially ordered combinations of syllables. Syllables are often considered the phonological "building blocks" of the words of a particular language. The syllables usually contain an onset, a nucleus and a coda. "Nucleus" is usually a vowel-like sound where 'onset' and 'coda' are consonant clusters.

We created 4 sets of corpuses for the recordings, each set containing 7 different script sets (one for each emotion category and one for the neutral category). The first corpus set was generated by replacing the entire vowel nuclei and consonant clusters in the selected Dutch texts using a weighted swapping mechanism in accordance with the natural probability distribution of the *vowel nuclei* and

the *consonant clusters* of Dutch. For the generation of the second corpus set, the entire consonant clusters in a Dutch text were replaced in accordance with the natural probability distribution of *voiceless consonant clusters* of Dutch while the vowel nuclei were replaced in accordance with the natural probability distribution of the *vowel nuclei* of Dutch. The third and the fourth corpuses were created accordingly but this time using English texts and the corresponding probability distributions of *vowel nuclei*, *consonant clusters* and *voiceless consonant clusters* of English. The structure of the four corpuses are summarized in Table 1.

The probabilities of occurrence in English and Dutch are calculated for each vowel nucleus (as explained in [4]) and for each consonant cluster. For consonant clusters begin (onset), middle and end (coda) consonant cluster probabilities were calculated separately. Similarly, the same calculations are performed for the voiceless consonant clusters (begin, middle, end). The probabilities were calculated using texts of approximately 27000 words from a large online text corpus - Project Gutenberg [5].

Table 1. *The summary of the corpus structures*

Corpuses			
NAME	LANGUAGE	CONSONANT DISTRIBUTION	VOWEL DISTRIBUTION
C1	Dutch	Whole consonant space	Whole vowel space
C2	Dutch	Voiceless consonant space	Whole vowel space
C3	English	Whole consonant space	Whole vowel space
C4	English	Voiceless consonant space	Whole vowel space

The texts were categorized in a way that we would have controlled variation in the sentences. These sentences contained different number of words, starting from one word up to ten words. In each emotion category the proportion of the number of words was the same.

The sentences were organized in paragraph structure to provide a dialogue impression. This is the kind of structure similar to dialogues used in theatre/film scripts.

2.3 Actual Recordings

Setup. The recordings took place in our recording lab [6] where the proper acoustic absorption was provided. The speaker was sitting on a stool chair with a proper headphone. The microphone (Neumann U87) was at a fixed position from the mouth of the speaker. Reading pane was put at a position where the speaker felt comfortable. Fig. 1 shows the recording set up.

The control room was outside the recording chamber and there was a window connecting the rooms visually.



Fig. 1. The recording setup.

Recording Procedure. The recordings started with voice tuning practices. The voice type should have suited the robotic character communicating with children. On the other hand, as the speaker would use the same type of the voice for a long period of time, it was important to find the voice type that the speaker felt comfortable with. We let the speaker improvise a few different voice types and recorded all of them. Considering the above two criteria, we chose one of the voice types in consultation with the speaker. During the recordings, we periodically played back the recorded sample of the voice type in order to keep the voice type stable during the entire recording session.

We repeated the same reference building procedure before each emotion recording as well. Taking the recorded base voice as a reference, the speaker improvised each emotion with that voice type. Then we kept the final sample as a reference for that emotion and let the actress train for a while. At the beginning of each script paragraph, we played the reference and the speaker continued acting in the same voice quality of the emotion. Also during the recordings, whenever a difference in the level/quality of emotion or voice type was noticed, that part was compared with the reference and re-recorded if needed.

A stuffed prototype of Probo was put in the recording room. This helped the speaker to act as being the robot. The photographic facial expressions of the robot were pinned on the face of the stuffed prototype to visualize the robot's emotions. The speaker found that method very helpful for getting back in the mood.

Before the recordings, a short discussion was held with the speaker about how to get in the mood for the different emotions. The speaker was also a drama trainer for children. She told us that in their acting trainings, they let the trainees close their eyes and relive some scenes from their lives that had the particular moods/emotions. We let her use the same method that she was used to to put herself in the mood. Only when she could not bring any scene from her life, we told her a short story in that particular emotion about Probo.

The speaker chose the emotion as well as the text corpus to start with. We planned 5-10 minutes of breaks hourly but the speaker could also take a break whenever she felt tired.

The recordings were done with Pro-Tools 8 and the pre-amplifier used was Earthworks 1021. All the data is recorded with 48 kHz sampling rate and 24 bits.

3 Evaluations

3.1 Experiments

We performed a series of two experiments; one with adult listeners and one with children listeners. While more subjects participated to the children experiment, the audio part of the children experiment was structured as a subset of the adults experiment. Only one database subset (C1) was used in the children experiment. Aside from the audio section, the children experiment has also included visual and audiovisual sections. The children experiment is analyzed and discussed in detail in [7].

Ten subjects participated to the adult experiment. The age of the subjects varied between 27 and 32.

Random samples were selected from each database subset (C1, C2, C3, C4) for each emotion category. The length of the samples had to be long enough so that the subjects could evaluate effectively. On the other hand, the length should not be too long not to lose the attention of the participants. So we decided to use 10 seconds of samples. Four different samples of 10 seconds are created for each emotion.

We instructed the subjects to listen to a number of samples of which they might not understand the meaning. The order of the samples were distributed randomly across emotions and we only used a single presentation order for all the subjects. The subjects were requested to choose which one of the possible emotions *anger*, *disgust*, *fear*, *happiness*, *sadness*, *surprise* or *neutral* matched the speech sample they heard. Subjects were allowed to listen to the samples as often as they desired.

As the final goal is to create a *natural sounding* gibberish language that can be used in building expressively interacting computing devices, the naturalness of the database had to be evaluated. Thus, in a second question, the subjects were asked to pay attention to the naturalness of the samples. They were instructed that the sample was considered as natural when it sounded rather like an unrecognized real language and not as an unnatural or random combination of sounds. Subjects were asked to assess their perception of the naturalness of the samples using Mean Opinion Scores (MOS) in a scale from 1 to 5. We also asked them to write down the language if the sample sounded like a language they knew to investigate if it is still possible to recognize the original language of the corpuses after consonant and vowel swapping.

3.2 Results

Fig. 2 shows the emotion recognition results for all 4 experimental corpuses (C1, C2, C3, C4). “Correct” stands for the emotion that was perceived as the intended

emotion and “incorrect” stands for the emotion that was perceived as one of the other emotions and not the intended one. As can be seen from the graphs, there is not a big difference in the recognition results which was also confirmed by the Kruskal-Wallis test.

When we analyzed the results emotion-by-emotion, a statistical significant difference is found only with *happiness* among the different corpuses. The recognition result of *happiness* was significantly lower in C2 than the other corpuses.

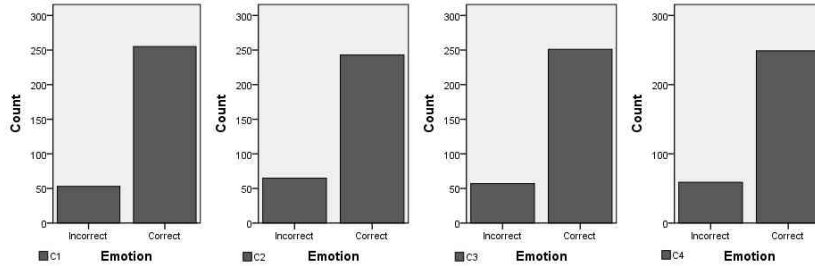


Fig. 2. Emotion recognition results for all 4 experimental corpuses (C1, C2, C3, C4, from left to right).

Overall/combined emotions versus recognized emotions are shown in the confusion matrix of Table 2. *Sadness* was recognized by most of the participants (94%). The recognition rate of *sadness* was followed by *neutral* with 88%, *surprise* with 87%, *happiness* with 84%, *disgust* with 74%, *fear* with 73% and *anger* with 66%. *Fear* was usually confused with *surprise* and *anger* was usually confused with *neutral* or *surprise*.

In the children experiment in which only C1 was used, *sadness* was recognized the best (100%). This was followed by *surprise* with 86%, *fear* with 71% and *disgust* with 57%. *Happiness* was often confused with *anger* and vice-versa which resulted in a lower recognition (29% and 46%, respectively). Much better results were achieved in the adult experiment for the same corpus C1 (91% and 64% for *happiness* and *anger*, respectively). This difference can be an indication that children and adults might have a different interpretation of, especially, *happiness*. For the other emotions, the recognition rates for C1 in the adult experiment were as following: 100% for *sadness* and *surprise*, 91% for *fear*, 55% for *disgust*.

Table 3 shows the average MOS scores for each corpus. As can be seen, the overall mean score is 3.6. This implies that the gibberish speech is perceived as natural by most of the subjects. The MOS results of corpus C1 was slightly higher than the other corpuses but a Kruskal-Wallis did not show a significant difference.

In the children experiment, the participants were provided with the question requesting a Mean Opinion Score (MOS) for the voice. The average MOS score for if the subjects liked the voice was 7.03 (out of 10).

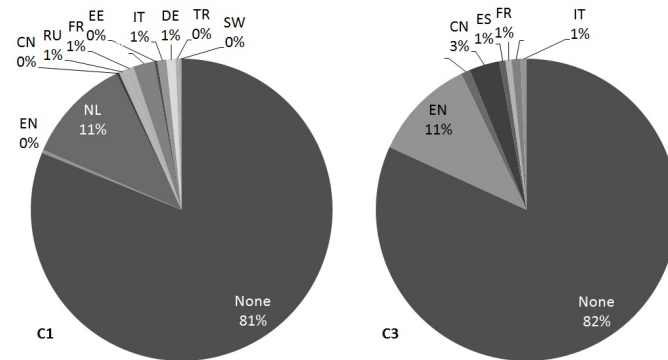
Table 2. Overall confusion matrix (expressed in %)

	Neutral	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Neutral	87.5	1.1	0.0	0.0	9.7	0.6	1.1
Anger	13.1	66.5	3.4	1.1	5.7	0.6	9.7
Disgust	4.5	8.0	75.0	0.6	2.3	8.0	1.7
Fear	0.0	6.3	0.0	73.3	1.1	7.4	11.9
Happiness	1.7	0.6	0.6	2.8	84.1	5.7	4.5
Sadness	0.6	1.1	0.0	4.0	0.6	93.8	0.0
Surprise	2.3	3.4	0.6	1.7	1.1	4.0	86.9

Table 3. Experimental results for MOS scores

Corpus	Mean MOS
C1	3.7
C2	3.6
C3	3.6
C4	3.5
General Mean	3.6

Fig. 3 shows to what extent the subjects were able to identify the original language in C1 and C3. It was seen that, for most of the subjects both of the corpuses did not sound as any language they knew. For the samples that the subjects thought they had recognized an existing language, the majority of them suspected these to be Dutch or English, for C1 and C3 respectively.

**Fig. 3.** Percentages of language recognition for C1 and C3 corpuses.

4 Conclusions and Further Work

In this paper, we described our emotional gibberish database with its primary aim of affective communication between robots and their children users.

The perception experiments showed respectable emotion recognition results of up to 81% overall (and even up to 94% for certain emotions). No statistically significant difference is found in the overall recognition results and emotion-wise (only with an exception of *happiness*) between all the four unique corpuses. This means that our methodology of recording induced emotions from an actor gave stable recognition results. We believe that the main driving reason for this stability was mostly our utilization of the control/reference sentence which was described in Section 2.3.

Between the children and the adult experiments, a remarkable difference was noticed with *happiness* (29% and 91%). This might be an indication that children and adults might have a different interpretation of *happiness* but further research is needed to check this hypothesis.

It is seen that the gibberish language we created resembles a natural language for most of the subjects (with an overall mean score of 3.6 on a scale of 1 to 5). That is important since our goal is to create a meaningless language that sounds like a real language.

In general, the gibberish language we created does not sound as any other languages known by the subjects. For the corpuses where a natural distribution of consonants and vowels was used (C1 and C3), the gibberish speech still sounded slightly like the languages of the texts that were used to create the gibberish texts.

As no statistically significant difference is found between the four different corpuses, for both emotion recognition results as well as for the naturalness, we can use all the four corpuses for emotional speech communication studies.

Combining the results from adult experiment that the gibberish speech resembled a natural language with an average MOS of 3.6 (out of 5) with the results of the children experiment that they liked the voice with an average MOS of 7.0, we can conclude that this database can be used in further studies focusing the children aged 10 to 14.

Apart from the described usage, our database could also be used as a segmental evaluation method for synthetic speech [8] or to test the effectiveness of affective prosodic strategies [9], and it can also be applied in actual systems [10], [11]. With the data recorded we also have a large interest to study the turn taking process of a conversation. As the text corpus was structured in a paragraph manner, start and stop sentences exist in the database. We envision that by analyzing the data recorded we will be able to develop a two-way conversation utterance structure for human robot interaction.

Acknowledgments. The research reported in this paper was supported in part by the Research counsel of the Vrije Universiteit Brussel with horizontale onderzoeksactie HOA16 and by the European Commission (EU-FP7 project

ALIZ-E, ICT-248116). Special thanks to Mr Ronny Van Heue, the principal and the students of Koninklijk Atheneum Beveren for their help and support in the experiments with the children.

References

1. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: Towards a new generation of databases. *Speech Communication*. vol 40, 33–60 (2003)
2. Busso, C., Narayanan, S.: Recording audio-visual emotional databases from actors: a closer look. In: *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation - LREC* (2008)
3. Wilson, D.E., Reeder, D.A.M.: *Mammal Species of the World: A Taxonomic and Geographic Reference*. Johns Hopkins University Press (2005)
4. Yilmazyildiz, S., Latacz, L., Mattheyses, W., Verhelst, W.: Expressive Gibberish Speech Synthesis for Affective Human-Computer Interaction. In: Sojka, P., Horak, A., Kopecek, A., Pala, K. (eds.) *TSD 2010. LNCS*, pp. 584–590. Springer, Heidelberg (2010)
5. Hart, M., Project Gutenberg, 2003, <http://www.gutenberg.org>
6. ETRO Audio-Visual Lab, http://www.etro.vub.ac.be/Research/Nosey_Elephant_Studios/
7. Yilmazyildiz, S., Henderickx, D., Vanderborght, B., Verhelst, W., Soetens, E., Lefeber, D.: Multi-Modal Emotion Expression for Affective Human-Robot Interaction (paper submitted)
8. Carlson, R., Granström, B., Nord, I.: Segmental Evaluation Using the Esprit/SAM Test Procedures and Mono-syllabic Words. In: Bailly, G., Benont, C. (eds.) *Talking Machines*, pp. 443–453 (1990)
9. Yilmazyildiz S., Mattheyses W., Patsis G., Verhelst W.: Expressive Speech Recognition and Synthesis as Enabling Technologies for Affective Robot-Child Communication. In: Zhuang, Y., Yang, S., Rui, Y., He, Q. (eds.) *PCM 2006. LNCS*, vol.426, pp. 1–8. Springer, Heidelberg (2006).
10. Oudeyer, P.Y.: The Synthesis of Cartoon Emotional Speech. In: *International Conference on Prosody*, pp. 551–554. Aix-en-Provence, France (2002)
11. Breazeal, C.: *Sociable Machines: Expressive Social Exchanges Between Humans and Robots*. PhD thesis, MIT AI Lab. (2000)
12. Saldien, J., Goris, K., Yilmazyildiz, S., Verhelst, W., Lefeber, D.: On the design of the huggable robot Probo. *Journal of Physical Agents, Special Issue on Human Interaction with Domestic Robots*. vol. 2. (2008)