

# Expressive Gibberish Speech Synthesis for Affective Human-Computer Interaction

Selma Yilmazyildiz, Lukas Latacz, Wesley Mattheyses, and Werner Verhelst

Interdisciplinary Institute for Broadband Technology - IBBT,  
Vrije Universiteit Brussel, dept. ETRO, Belgium  
{syilmazy, llatacz, wmatthey, wverhels}@etro.vub.ac.be  
<http://www.ibbt.be/>, <http://www.etro.vub.ac.be/>

**Abstract.** In this paper we present our study on expressive gibberish speech synthesis as a means for affective communication between computing devices, such as a robot or an avatar, and their users. Gibberish speech consists of vocalizations of meaningless strings of speech sounds and is sometimes used by performing artists to express intended (and often exaggerated) emotions and affect, such as anger and surprise, without actually pronouncing any understandable word. The advantage of gibberish in affective computing lies with the fact that no understandable text has to be pronounced and that only affect is conveyed. This can be used to test the effectiveness of affective prosodic strategies, for example, but it can also be applied in actual systems.

**Key words:** Affective Speech Synthesis, Expressive Speech Synthesis, Gibberish Speech

## 1 Introduction

The desire of mankind for an intelligent interaction with machines (HCI) has been among the mostly dreamed of concepts in science fiction and exact science alike. Today, humans communicate with machines in everyday life. Navigation systems, car diagnosis systems, computer games, distance learning applications, assistive technologies and robotic assistants are just a few examples. Day by day, HCI resembles more and more the natural interaction between humans (HHI). One of the most important interaction features of HHI is expressive speech communication, which allows the communication of affect and intent and this not only between humans but also with animals, as used in animal assisted therapy [1] and robot assisted therapy [2].

In HHI, only 7% of information is transferred by the words spoken while 38% is transferred by the tone of voice [3]. Therefore, a nonsense language like gibberish could be successful as a carrier to express emotions and affect. Gibberish could even be more advantageous than plain speech since no understandable text has to be pronounced and the focus is only on conveyed affect. Our aim is to contribute to affective HCI a non-understandable/gibberish language and to

build expressively interacting computing devices. We also intend to experiment with affective gibberish speech for communication between robots and children.

The paper is organized as follows: in section 2 we describe our approach for gibberish text generation; in section 3 we discuss its usage as a front end for text to speech synthesizers (TTS); in section 4 we investigate the correlations between perceived and intended emotions in both plain and gibberish speech and in section 5 we conclude with a discussion.

## 2 Gibberish Speech Synthesis

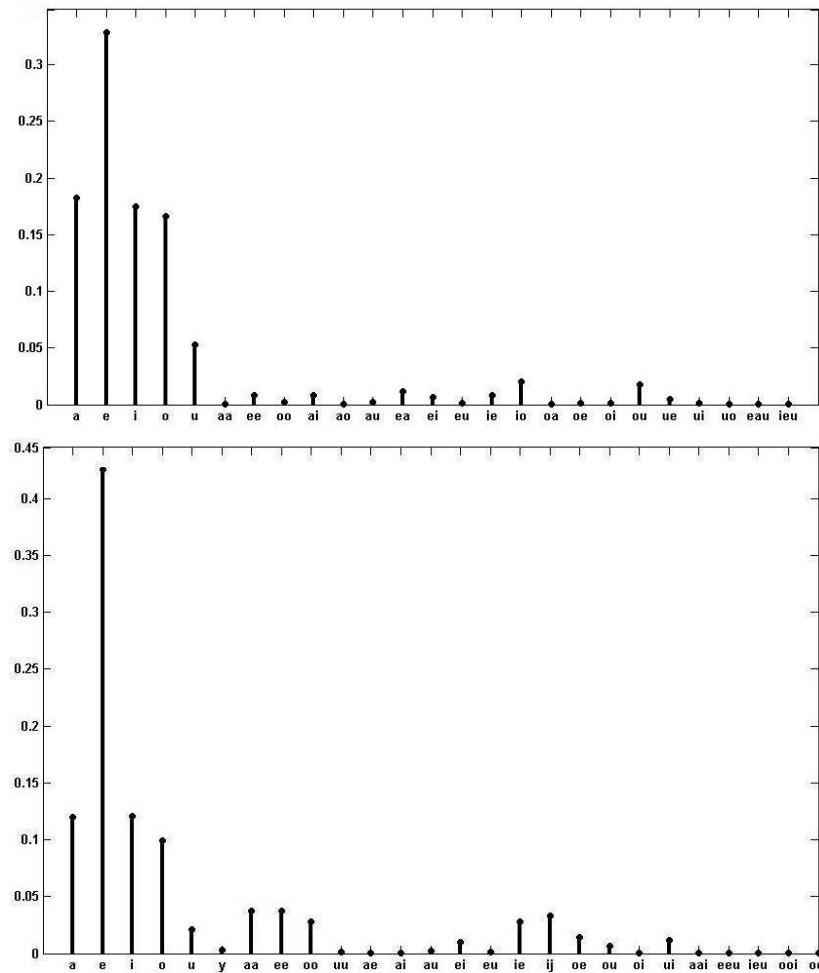
Siblings sometimes use toy language. Such language can be either a coded form of their mother tongue (e.g., “mother tongue” becomes “mopotheper topongue” in the p-language) or can be meaningless (i.e., gibberish). Meaningless speech can also be used as a segmental evaluation method for synthetic speech [4], to test the effectiveness of affective prosodic strategies [5], for example, but it can also be applied in actual systems [6],[7].

Languages consist of ruled combinations of words and words consist of specially ordered syllables. Syllables are often considered the phonological “building blocks” of the words and they usually contain an “onset”, a “nucleus” and a “coda”. In most languages, every syllable requires a nucleus which is usually a vowel-like sound. In English and Dutch, vowel nuclei are transcribed with one, two and three letters. There are usually only a few vowel nuclei with one letter transcriptions but they are most frequently used in the language (fig. 1). There are usually much more vowel nuclei with two or three letter transcriptions, but these are far more rarely used.

To produce gibberish speech, we wrote 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 that gibberish text as input for TTS engines to generate the gibberish speech. However, if we would transform the word “language” into gibberish with a uniform random swapping of vowel nuclei, we would likely end up with something like “lieungeaugie”. To avoid this, we calculated the probabilities of occurrence for each vowel nucleus and used a weighted swapping mechanism in accordance with the probabilities instead of uniform random swapping. Fig. 1 represents the empirical probability distributions of vowel nuclei for English and Dutch text of approximately 27000 words each from Project Gutenberg [8].

## 3 Input Text And Language

Our goal being to create a *natural sounding* gibberish language, we transform existing text into meaningless text and use this as input text for a TTS. However, as the TTS’s language processing modules are not designed to work on meaningless text we investigated how natural our synthetic gibberish sounds and whether the native language of the TTS affects the result.



**Fig. 1.** Empirical probability mass distribution of vowel nuclei in English (upper panel) and Dutch (lower panel).

We therefore created two sets of sentences. For the first set, 6 original English sentences were selected from children’s stories [8] and converted to gibberish using the English vowel nuclei probability distributions. For the second set, 6 sentences were selected from Dutch children’s stories [8] and converted to gibberish using the Dutch vowel nuclei probability distributions. We then synthesized all 12 gibberish sentences both with the Dutch and the English version of our unit selection TTS [9]. We thus constructed 4 different groups of samples: 6 samples with Dutch gibberish text and Dutch TTS, 6 samples with English gibberish text and English TTS, 6 samples with Dutch gibberish text and English TTS, and 6 samples with English gibberish text and Dutch TTS.

Ten subjects aged between 24 and 37 participated in a listening experiment. Four subjects had no prior experience with synthetic speech. The subjects were asked to pay attention to the naturalness of the samples. They were instructed that a sample is to be considered as natural when it sounds more like an unrecognized real language than like an unnatural or random combination of sounds. They were asked to express their judgement using Mean Opinion Scores (MOS) on a scale of 1 to 5. We also asked them to write down the language if the sample sounded like a language they knew. For the naïve subjects, we provided an example of natural (i.e., plain) synthetic speech at the beginning of the test to ensure that they would not rate the quality of the TTS instead of the naturalness of the gibberish.

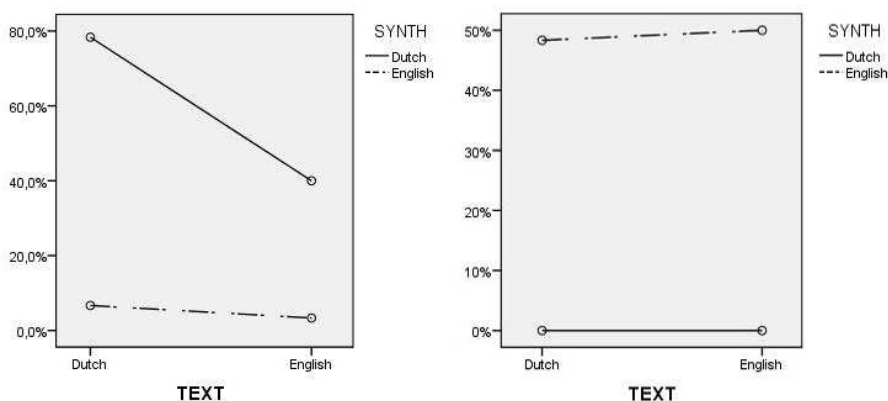
Table 1 shows the average MOS and the results of the 2x2 ANOVA on the MOS scores. The two variables are the original language of the input gibberish text (TEXT) and the language used for synthesis (SYNTH). The gibberish speech was perceived as natural by most of the subjects with an overall MOS of 3.62. The language of the synthesizer had a significant influence (Sig. = 0.043) on the perceived naturalness but no significant influence of the input language and no combined effect were found. The samples created with the Dutch synthesizer had the highest score for both Dutch and English gibberish texts. That could be because almost half of the subjects were native Dutch speakers. Informal listening has shown that the Dutch synthesizer has better quality than the English version, which could be another possible explanation. In general, we can conclude that all versions of the gibberish speech synthesizer were found to be rather natural sounding.

Fig. 2 shows to what extent the subjects were able to identify the original language in the gibberish samples. It is seen that for both Dutch and English, the recognition rates are highest when both the gibberish input text language and the synthesizer language are the same. In general the synthesizer language has more impact while the text language has only little effect. Dutch was more easy to recognize with scores up to 78% while for English the highest recognition rate was about 50%. This is most likely due to the fact that almost half of the subjects were native Dutch speakers and that the Dutch synthesizer has better quality than the English version. An important conclusion is that gibberish speech with the correct probability distribution of Dutch vowel nuclei and synthesized with a Dutch TTS system does indeed resemble Dutch. Together

**Table 1.** *Experimental results and statistical analysis (sig. threshold level  $\alpha=0.05$ )*

Test Results			2x2 ANOVA			
TEXT	SYNTH	Mean MOS	Factor	df	F	Sign.
Dutch	Dutch	3.82	TEXT	1	0.000	1.000
Dutch	English	3.42	SYNTH	1	4.137	0.043
English	Dutch	3.68	TEXT*SYNTH	1	1.034	0.310
English	English	3.55				
General Mean		3.62				

with the results presented in Table 1, we may thus conclude that we achieved our goal of constructing a gibberish synthesizer that sounds like natural Dutch.

**Fig. 2.** Percentages of language recognition for Dutch (left panel) and English (right panel).

## 4 Semantic Meaning And Perceived Emotion

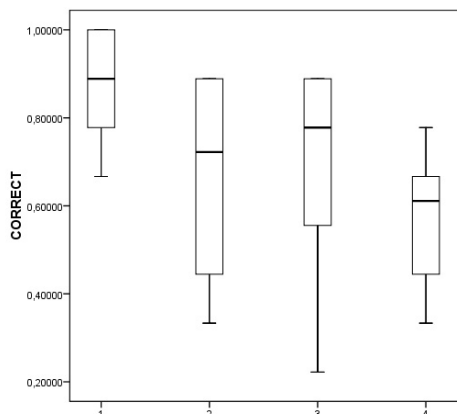
People naturally use both prosodic meaning and semantic meaning for expressing affect and emotion. In gibberish, there is no semantic information. Furthermore, the fact that gibberish is meaningless might interfere with the prosodic strategy of the synthesizer and result in less expressive speech. Therefore, we investigated whether the semantics of the underlying text influence the perception of emotions in synthetic speech and whether gibberish might be more or less effective than plain speech in conveying the intended emotion.

We synthesized 4 groups of samples. In the first group, the semantic meanings of the sentences and the acoustic properties of the synthesized utterances corre-

spond to the same emotion. In the second group, the semantic meanings of the sentences have opposite emotion of the acoustic properties. In the third group, the semantic meanings of the sentences are neutral, and in the fourth group the sentences are gibberish and therefore have no semantic meaning. Each group contains 10 samples. The emotion categories used were happy and sad. We used the open source emotional TTS synthesis tool, “EmoSpeak”, of the synthesizer Mary [10],[11] with the parameter settings for happy and sad reported in [12] to produce the emotional speech.

Nine subjects aged between 26 and 37 joined the forced-choice listening test. Three subjects had no experience with synthetic speech. The subject were instructed to listen to a number of samples of which they may or may not understand the meaning and they were requested to choose which one of the possible emotions happy, sad or neutral matches the sample they heard.

Fig. 3 shows the emotion recognition results for all 4 groups. Group 1 (semantic meaning and acoustics correspond to the same emotion) has the highest scores among all groups. The other 3 groups showed comparable recognition results amongst each other. A Kruskal-Wallis test confirmed that Group 1 is indeed significantly different (Sig.=0.032). Thus, semantic meaning did help for recognizing the intended emotion, as expected. On the other hand, semantics opposite to the intended emotion did not make the task more difficult than with neutral semantics or with gibberish speech.



**Fig. 3.** Box plot of the emotion recognition results for 4 different experimental groups.

## 5 Discussion

In the first experiments we explored the influence of input text and language on the synthesized gibberish speech. It was found that gibberish speech resembles

natural language with a total average MOS of 3.62. We did find a significant difference between naïve subjects and subjects having expertise with synthetic speech. The overall naturalness ratings of the naïve subjects were significantly lower than the ratings of the speech experts. We received feedback from the naïve subjects that they found it difficult to evaluate the naturalness of gibberish speech independent from the synthesis quality; they believe that the synthetic speech quality may have negatively influenced their scores. Also, the samples from the Dutch synthesizer received higher scores than the samples from its English version, which is likely due to the difference in quality between the Dutch and English versions of the synthesizer as the Dutch synthesis database is almost 4 times larger than the English database.

The experiments also showed that the gibberish language resembles the source language when a good quality synthesis is used in combination with an input text from the same language. This can be easily understood as, even without semantic meaning, the synthesizer still uses the phones and intonation model of its target language. Moreover, the gibberish input text for the synthesizer has the same vowel distribution as the source language.

From the experiments on the relation between semantic meaning and the perceived emotion, we found that semantics help for recognizing the intended emotions when the semantic and the prosodic meaning of the utterances are both in line with the intended emotion. When they were in line with opposite emotions, this did confuse the subjects but less so than might have been expected. A probable cause is that the synthesizer we used simulates happy with high speaking rate and sad with low speaking rate such that the intended emotion could be easily inferred. We received feedback from subjects that they did indeed mostly use speaking rate as a clue to infer the intended emotion.

No statistical difference was found between samples with emotionally neutral meaning and gibberish samples. As a consequence, we can say that gibberish speech conveys the emotions as effectively as semantically neutral speech and can be used in an affective communication system. It should be noted, however, that this conclusion is valid for our experiments on emotional speech synthesis with the Mary synthesizer and it is not at present clear to what extent similar results would be obtained with other expressive synthesizers.

**Acknowledgments.** The research reported on in this paper was supported in part by the Research counsel of the Vrije Universiteit Brussel with horizontale onderzoeksactie HOA16, by the Flemish government (IBBT project SEGA) and by the European Commission (EU-FP7 project ALIZ-E).

## References

1. Heimlich, K.: Animal-assisted Therapy and the Severely Disabled Child: A Quantitative Study. *Journal of Rehabilitation*, vol.67, no.4, pp. 48–54, October/November/December (2001)

2. Shibata, T., Mitsui, T., Wada, K., Touda, A., Kumasaka, T., Tagami, K., and Tanie, K.: Mental Commit Robot and Its Application to Therapy of Children. In: Proc. of the IEEE/ASME International Conference on AIM'01, pp. 182, July (2001)
3. Mehrabian, A.:Communication Without Words. *Psychology Today*, vol.2, no.4, pp. 53–56 (1968)
4. Carlson, R., Granstrm, B., Nord, I.: Segmental Evaluation Using the Esprit/SAM Test Procedures and Mono-syllabic Words. In: *Talking Machines* (G. Bailly, C. Benont eds) Elsevier, North Holland, pp. 443-453 (1990)
5. 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 Berlin, Heidelberg (2006)
6. Oudeyer, P.Y.: The Synthesis of Cartoon Emotional Speech. In: Proc. of the 1st International Conference on Prosody, pp. 551–554. Aix-en-Provence, France (2002)
7. Breazal, C.: *Sociable Machines: Expressive Social Exchanges Between Humans and Robots*. PhD thesis, MIT AI Lab. (2000)
8. Hart, M., Project Gutenberg, 2003, <http://www.gutenberg.org>
9. Latacz, L., Kong, Y., Mattheyses, W., Verhelst, W.: An Overview of the VUB Entry for the 2008 Blizzard Challenge. *Blizzard Challenge 2008*, Brisbane, Australia (2008)
10. Schröder, M.: *Speech and Emotion Research: An Overview of Research Frameworks and Dimensional Approach to Emotional Speech Synthesis*. PhD thesis, PHONUS 7, Research Report of the Institute of Phonetics, Saarland University (2004)
11. OpenMary: Open Source Emotional Text-to-Speech Synthesis System, <http://mary.dfki.de/>
12. Schröder, M., Cowie, R., Douglas-Cowie, E., Westerdijk, M., Gielen, S.: Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis. In: Proc. of the Eurospeech 2001, vol. 1, pp. 87-90, Aalborg (2001)