Understanding Factors in Emotion Perception

Lakshmi Saheer, Blaise Potard

Idiap Research Institute, Martigny

lsaheer@idiap.ch, bpotard@idiap.ch

Abstract

Emotion in speech is an important and challenging research area. Addition or understanding of emotions from speech is challenging. But, an equally difficult task is to identify the intended emotion from an audio or speech. Understanding emotions is important not only in itself as a research area, but also, for adding emotions to synthesized speech. Evaluating synthesized speech with emotions can be simplified if the correct factors in emotion perception can be first identified. To this end, this work explores various factors that could influence the perception of emotions. These factors include semantic information of the text, contextual information, language understanding and knowledge. This work also investigates the right framework for a subjective perceptual evaluation by providing different options to the listeners and checking which ones among these are the most effective response to evaluate the perception of the emotion.

Index Terms: Emotion perception, perceptual factors, emotions in speech, emotional speech analysis/synthesis, metrics for subjective perceptual evaluations

1. Introduction

In recent years, many researchers have been studying the area of emotional speech, how to synthesize, recognize or even interpret emotions from human speech. Emotion in speech is a rather complex topic compared to other research areas in speech, particularly due to the absence of a standard metric to gauge the emotional content of a speech sample.

By contrast, automatic speech recognition (ASR) can be evaluated in a simple manner using objective scores such as word or phone error rates. Text-to-speech synthesis, while harder to evaluate parametrically, still has objective and subjective scores related to quality or intelligibility. There are no such parameters to classify the emotions expressed in speech. Listeners usually identify the emotions subjectively, according to their mood, opinions and cultural background. As a result, a given sample of emotional speech could be perceived differently by each listener.

There are various techniques used to generate the emotion in synthesized speech. A short review of the different emotion generation techniques was presented in [1]. There are 3 major techniques explained in [1] for addition of emotions: (1) Formant Synthesis: the acoustic speech data is entirely generated using the rules on acoustic correlates of various speech sounds. Emotional expressivity is modeled by manipulating the parameters related to voice source and vocal tract. (2) Di-phone synthesis: the monotonous human recordings are split into di-phones and then concatenated during synthesis. The required F0 contour is generated through signal processing techniques, which along with some manipulation of the duration can generate the emotion expression to some extent. (3) Unit selection synthesis: units of variable size are selected from a large inventory of speech database which best approximates a desired target utterance defined by a number of parameters. Databases for different emotions are collected separately and corresponding units are concatenated for generating emotional speech. Apart from the techniques mentioned above, there are also efforts with parametric speech synthesis [2] to generate different speaking styles and expressions. It is easier to generate different emotions and speaking styles with a hidden Markov model (HMM) based parametric speech synthesis system by using a style vector as in [3]. A comparison of unit selection and HMM based emotional speech synthesis [4] revealed that unit selection methods require improvements to prosodic modeling and that HMM-based methods require improvements to spectral modeling for emotional speech synthesis and that certain emotions cannot be reproduced well by either method.

The accuracy of humans listeners for the emotion identification task for recorded speech databases is approximately 80% [5]. This number will further degrade if a machine generated emotion or speech is being evaluated. The degrees of perceived emotions also varies from speaker to speaker and across listeners. Various techniques have been proposed to evaluate emotion - both objectively and subjectively [6]. The objective scores include classification of speech based on spectral or prosodic features for emotion identification using machine learning algorithms. Subjective tests are more popular with emotion evaluation since emotions are supposed to be perceived more realistically by human listeners. The most popular subjective tests include forced categorization of emotions into prespecified classes, or a descriptive free response system. The evaluation task can be made more intricate by assessing the degrees of naturalness, believability or overall preference of the emotion expression (often on a five-point scale) in addition to the forced categorization [7].

There are different perspectives to interpret emotions, in particular a psychological and a signal processing perspectives. The psychological perspective consists of studying different states or degrees [8] (e.g. activation, valence, dominance) to differentiate emotions. The signal processing perspective studies the changes in the speech signal attributes corresponding to different emotions [9]. There are also different ways of evaluating the emotion perception as summarized in [1] which represents the "perceptual" view for understanding emotions. This study aims at collecting these different streams of emotional research under one roof with the aim of studying how the different emotion generation techniques can be effectively evaluated in future. Towards this end, this work performs perceptual evaluations for different emotions with/without corresponding semantic and contextual information. The influence of language knowledge and understanding is also evaluated for the perception of different emotions. Finally, the best mode of response from listeners is evaluated among the three response types pro-

posed: emotion classification using a forced choice task, emotion identification as a free text response and identifying the degrees/states for different emotions.

This paper is organized as follows. Section 2 briefly describes the factors that could influence the perception of emotion, followed by the design of the subjective listening tests explained in section 3. Finally, the deductions from the evaluations are summarized in section 4.

2. Factors affecting emotion evaluation

There are various factors that affect the perception of emotion, including the mood and perception of the listeners. Studying an exhaustive list of all such factors would greatly exceed the scope of this paper. In this work, we focus on four basic factors that could influence the emotion perception: (1) Parameters of the speech signals that indicate the particular emotion; the basic features include pitch, duration and intensity, but there are numerous other voice quality features as well. (2) Semantic content of the text that is being spoken; we are interested in studying whether people can perceive the emotion if the semantic content of the text does not match the emotion that is being expressed. This could in turn lead to the topic of representing irony[1] in synthesized speech. (3) Context Information which is intended to help understand the mood or situation of the speaker. (4) The language ability of the listener could influence the perception of emotions, as it directly influences the understanding the semantics of the emotional speech samples.

This work hypothesizes that these four factors influence the perception of emotion, and the authors expect the results of the evaluations to reveal whether the hypothesis can be proved. Since it is not possible to gauge mood of the listeners or the variation in emotion perception by different listeners, these factors are usually (assumed to be) statistically normalized using a large enough number of listeners (of the order > 10). The different factors evaluated are detailed below.

2.1. Speech parameters

The emotional speech researchers have always relied on specific speech parameters that vary when representing a particular emotion. These features include pitch, duration, intensity, as well as some other voice quality features [10]. These features could be manipulated to modify the emotion in speech; the same features could be analyzed to perceive the intended emotion, which could lead to the generation of objective scores for emotion recognition. It is not clear however if human listeners use these cues in speech to perceive emotion. In order to understand if these cues play an important role in emotion perception, this work evaluates original recordings of emotional speech acted out by human actors. Compared to the evaluation of synthetic signals, this eliminates any direct effect from the distortions introduced by speech synthesis technology or emotion manipulation techniques. This way, the listener is not evaluating any particular emotion generation technique, but the intended emotions as expressed by humans.

2.2. Semantic content

A spoken utterance has two main components, the speech signal and the text that is spoken in the speech. The semantic content of the text cannot be isolated from the speech perception [11]. This factor might have a significant influence in the perception of emotion. Even though the listeners are usually asked to ignore the meaning of the text represented by the speech, it is not easy to totally isolate this influence. Most of the evaluations try to use "emotionally neutral" text. Neutral text itself can be classified into two types: one that fits into multiple emotions depending on the context - for example, the phrase "I did not expect this", could be expressed as happy, sad, or angry depending on the context in which it is spoken; the second type consists of neutral text that does not represent any particular emotion - for example, the phrase, "Whales live in the sea" does not convey any specific emotion. By opposition, emotive text is clearly intended to convey a specific emotion; an example of an emotive sentence is "I won a big lottery" and is clearly an "excited" or "happy" sentence. The emotions can be evaluated using sentences that represent neutral or emotive text (corresponding to the emotion). Similar studies were performed earlier with synthesized emotional speech [9]. This study further emphasizes the influence of the meaning of a sentence in perceiving the intended emotion.

2.2.1. Irony Effect

The influence of meaning can be further developed into the study of what is called the "irony effect". The irony effect refers to the situation where the speech express the opposite of (or something other than) the literal meaning of the sentence [12]. Other studies[19] attempt to synthesize the irony effect in machine-generated speech, by using an emotion different from the intended emotion in the text representing the speech. In this study, the irony effect is evaluated by using emotional speech for emotive sentences with different emotions, including ones that are opposite to the intended emotion in the text.

2.3. Context information

Most of the studies on emotional speech is based on a single sentence or phrase. Listeners are asked to evaluate the emotion on the basis of the perception from this sentence. This is a difficult task if the intended text is neutral or does not represent the whole mood of the speaker. The context in which the speech was spoken is particularly important for a listener to perceive the emotion of the speaker and the intended emotion in the sentence. This is similar to a real life scenario where the emotions are in a continuum and a particular emotion is expressed strongly in very rare situations. People usually understand the emotion of the speaker from the whole context and not from listening to ad-hoc sampled sentences. This study presents the listener with a detailed context in which the sentence is spoken and checks if the intended emotions are correctly identified in this case. Similar methodology was also tried in [9]. This method will recreate the real life scenario of how emotions are perceived by humans. The context information can be provided as an audiovisual content or even as a dialogue system. In this work, it is presented as a machine-generated background audio story with a few dialogues, which is efficient enough to convey the sentence context.

2.4. Language ability

A major factor affecting the emotion perception is the cultural impact. Different cultures represent emotions differently and the intended emotion might be totally different in different cultures even with the same speech parameters. There has been some study in this area [13]. An important factor to consider

[1] generally defined as *an expression or utterance marked by a deliberate contrast between apparent and intended meaning*[18]

is that the language ability might not only bias the listener towards the semantic meaning of the text, but also influences the emotion perception based on the cultural understanding of the language. To this end, emotional speech in a foreign language (German) is evaluated with some listeners who understand and others who have no knowledge of the language. In order to restrict the length of the evaluation, only one foreign language apart from English is evaluated.

3. Experimental Design

Recorded emotions of human actors are evaluated in this work, by opposition to machine-generated speech, so as not be subjected to the undesirable artifacts generally associated with synthesized emotional speech. This section gives the details of the design of the subjective evaluations performed to study the perception of the emotions based on the factors mentioned in the earlier section. The test is divided into four sections to evaluate the four factors mentioned above.

The first section represents the most commonly used experimental setup for emotion evaluation. The semantically neutral sentences are evaluated with different emotions in speech. The data used in this section comes from the EMA database [14]. Three semantically neutral sentences spoken by a male and a female speaker have been selected for the evaluation. Each sentence was spoken with four different emotions, namely happy, sad, angry and neutral, resulting in 24 sentences to be evaluated in this section.

The second section is similar to the first one but with speech from emotive text. Three different emotive sentences (with happy, sad and angry textual emotions) spoken by the same male and female speaker each with four different emotions (happy sad, angry and neutral) resulting in the evaluation of 24 sentences in this section. This section will help us study the influence of the semantic content of the speech signal. Since the same emotive sentence is used with different intended emotions in speech, this section may also give some insight on the perception of irony.

The third section represents the test for the influence of the contextual information in the perception of emotion. The same three emotive sentences used in the section 2 were presented within a context. Only the emotions corresponding to the emotive text was used for the test for both male and female speaker, which resulted in 6 sentences to be evaluated in this section. A state of the art commercial text-to-speech system[17] was used to generate the story around the emotional speech sample, to give a complete picture of the emotional state of the speaker. A voice very close to the speaker speaking in a neutral tonality was used to generate the context information, and the listeners were asked to ignore any distortion in the machine generated voice and evaluate the perception of the emotion from the single sentence highlighted as spoken by the human actor within the context of a neutral machine generated speech.

The last section is to study the influence of the knowledge of the language on the perception of emotion. The hypothesis to be tested here is that the knowledge of the language or the culture could improve the perception of the emotion. The tests were performed with emotional speech acted by a male speaker based on neutral text in German language from the German EmoDB [15]. Three sentences spoken with four different emotions including happy, sad, angry and neutral resulted in 12 sentences to be evaluated. Having listeners who do not understand any German may also allow us to study the perception of emotions without any semantic knowledge.

The responses of the listeners can be collected according to different response types as explained in [1]. The usual method of emotion evaluation is a discriminative task in which listeners are forced to select a particular emotion from a list of available emotion types. This, as mentioned in [1], is a discriminative task and not an emotion identification or descriptive task. This type of response makes the task simple and easy to evaluate. This task can be improved by adding a number of "distractor" response categories of emotions introduced in the perception test [9]. Other researchers [9] ask the listeners to describe in their own words what emotion they perceive. The listeners could provide a free response with keywords. These keywords are then grouped and classified into meaningful categories to identify the emotion perceived by the listener.

There are also ways to parameterize the emotions on different scales [8] based on states representing the emotions. This is similar to the scale represented by the FEELTRACE [16] concept representing emotions in a continuum between the space of valence and activation. This paper categorizes the emotional states as explained in [8] (also as conceptualized by the psychologists for the communication of affect), into three dimensions. These dimensions are usually termed as arousal, pleasure and power. Pairs of adjectives like happy/unhappy or pleased/annoyed (for pleasure), agitated/calm or excited/apathetic (for arousal) and powerful/powerless or dominant/submissive (for power) can be used to represent these three dimensions.

The types of responses in the evaluations for this work include a forced choice discriminative task with the four emotions actually used (happiness, sadness, neutral, anger) plus two distractor emotions (fear and surprise). Fear and surprise could be easily confused with sad and happy respectively. Also, the listeners are asked to provide a free response based on their perception of the emotion. A few listeners did not find this free response option useful and left the input space blank or mentioned "same as selection" or mentioned the same emotion they selected in the forced selection. This evaluation then presented three dimensions to evaluate or classify different emotions. These include valence, arousal/activation and dominance. Each of the three dimensions were varied on a five point Likert scale: Valence ranging from annoyed (negative) to pleased (positive), Activation or arousal ranging from very calm to very agitated/excited and Dominance varying from powerless (submissive) to powerful (dominant). The mid point of each dimension referred to as neutral may in turn represent a non-emotional speech. Apart from these responses, the "emotive text" section (section 2) also included a response to check if the emotional speech was perceived as irony or not. There were three choices: yes, no and maybe.

The test was available online, and was sent out to members of the research communities, mainly working on speech and signal processing. 16 listeners participated in the test, out of which 3 were native German speakers. The listeners were from different nationalities and cultures. Seven listeners did not understand any German while others varied from "can read/write" to "can understand bits". The test was rather long due to the multiple modes of responses and listeners found it a bit tiring especially the descriptive part of explaining the emotion in words.

4. Results and Discussions

Results of the perceptual experiments can be summarized with confusion matrices for the four sections separately. The confusion matrices are based on the forced choice test, which in-

Emotion	Keywords
Angry	quite excited, a bit annoyed, agitated, upset, disgust, panic, aggressive, pissed off, distressed, energetic, dominating, threatening, menace, stressed, disagreeing, jeering, unsatisfied, upset, declarative, impatient, strong, frustration, threat
Happy	elevated, enthusiastic, pleased, glad, ecstatic, excited, laughing, amused, having fun, optimistic, playful, joking, interested, content, optimistic, looking forward, confident, interested, positive emotion
Neutral	no emotion,disgusted, explanatory, declarative, determined, dubitative, unsure, calm, dubitative, questioning, tired, interested
Sad	upset, slightly disgust, grief, depressed, bored, confused, broken, beaten down, stoned, monotone, crying, tired, whispering, disappointed, weary, desperate, bored, empathic, melancholic, uninterested, nervous

Table 1: Keywords for emotion classes

cluded the two distractor emotions "fear" and "surprise". The descriptive response almost always had the same meaning as the forced choice representation and some listeners felt it as a redundant response that could be ignored. The keywords corresponding to the different classes are grouped together in Table 1. In a couple of isolated cases, listeners described the emotion correctly, but chose a different (wrong) emotion from the forced choice, but there were only very few cases that thus benefited from the descriptive response. This limited gain gets neutralized by other confusing keywords which do not have a clear class like the keyword "upset" which could represent either sad or angry. For these cases, the values given to the dimensions might help disambiguate the situation. To summarize, listeners preferred the forced choice due to convenience, and the descriptive response did not give any significant performance gain. The different degrees for the three dimensions were in good agreement across users.

The section evaluating the influence of the contextual information (section 3) lead to very accurate emotion classification. This task is however slightly confusing due to the presence of audio from two different origins in the stimuli: neutral machine-generated context speech, and emotional speech spoken by a human actor; the listeners were instructed to base their evaluations on the emotional speech. The listeners managed to correctly perform the evaluation task except for one listener who classified all data in this section as neutral, due to the neutral machine-generated contextual sentences. This listener has thus been omitted from the results of section 3. The results and corresponding observations for each section of the test are detailed below.

4.1. Section 1: Neutral Text

This section uses emotionally neutral text to evaluate the perception of emotion in speech. This is the classic way of evaluating emotional speech with the argument that when the text is semantically neutral listeners tend to focus on speech parameters conveying emotion. The results of the forced choice emotion discrimination is presented in Table 2. The table shows confusion matrix for the identified emotion with the underlying intended emotion in the speech. Angry and neutral have the best performance as observed in the literature, followed by the sad emotion. Happy emotion has the worst performance and is confused with surprise and neutral. The distractor emotion (surprise) here increases the confusion, degrading the performance. The table shows both recall and precision for representing the performance. All the performances are well above chance of 25%. The precision values are quite high even for the emotions that have a smaller recall. It is not clear whether listeners are ignoring the semantic meaning of the text and concentrating on the speech parameters since happy emotion is confused mainly with neutral, and neutral emotion has the best recall.

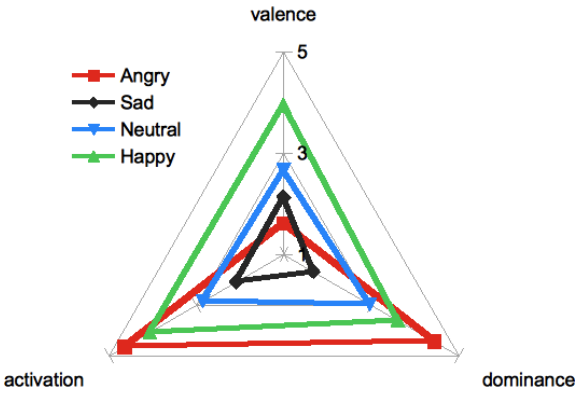

Figure 1: Average values for dimensions

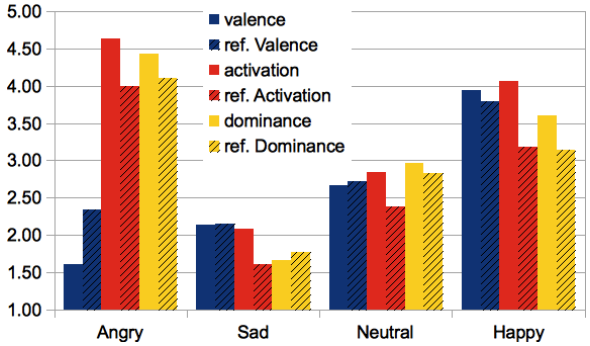

Figure 2: Comparison of Dimensions

In some cases, the neutral text with a particular emotion may be perceived emotive and hence, wrongly recognized (happy as surprise). This gives neutral emotion poor precision even with a good recall.

For the descriptive task, the keywords mostly correspond to the forced choice emotion selected. There are a few responses where the emotion is described correctly, but the forced choice emotion is wrongly selected. For example, selecting neutral in the forced choice task, but correctly describing as "tired or disappointment" for sad, "dominating" for angry, describing "happy" for happy (but selecting surprise) and "excited" for angry (selecting happy/surprise). Also, there are some responses that correctly identify the emotion in the forced choice task and describe them as a different emotion. For example, describing the speech as "annoyed", "depressed" or "angry" after correctly selecting the neutral and describing as "panic" after correctly choosing angry emotion in the discriminative task.

The different dimensions of the emotions are plotted in Figure 1 and also compared with what was reported with the database in Figure 2. The results of human evaluations performed by 18 listeners were distributed originally along with the database. Both results are comparable and follow the same trend across different emotions. The dimensions show clear regions of influence for each emotion with neutral emotion concentrating on the middle (at zero values). As seen in literature, the happy emotion overlaps with others in these dimensions and is not easy to differentiate.

4.2. Section 2: Emotive Text

The performance of the emotion identification improves for angry and sad with emotive text (as seen in Table 2), and performance of neutral and happy degrades. The degradation of neutral might indicate that the perception is based on the emo-

Sections	Neutral Text				Emotive Text				Context Info.		
Emotions	**A**	**S**	**N**	**H**	**A**	**S**	**N**	**H**	**A**	**S**	**H**
Angry	**75**	0	11	2	**90**	4	21	7	**29**	0	1
Surprise	8	0	0	23	1	0	0	25	0	0	2
Sad	1	**71**	7	5	1	**75**	16	0	0	**29**	0
Neutral	7	14	**77**	14	4	11	**59**	22	1	1	2
Happy	3	0	0	**51**	0	1	0	**41**	0	0	**25**
Fear	2	11	1	1	0	5	0	1	0	0	0
Recall	78.1%	74.0%	80.2%	53.1%	93.8%	78.1%	61.5%	42.7%	96.7%	96.7%	83.3%
Precision	85.2%	84.5%	68.8%	94.4%	73.8%	81.5%	61.5%	97.6%	96.7%	100%	100%

Table 2: Confusion matrix for forced choice test for first three sections

Dimensions	Valence				Activation				Dominance			
Emotions	**A**	**S**	**N**	**H**	**A**	**S**	**N**	**H**	**A**	**S**	**N**	**H**
Angry	-	0.004	0.004	0.002	-	0.002	0.002	0.007	-	0.002	0.002	0.004
Sad	0.004	-	0.007	0.002	0.002	-	0.004	0.002	0.002	-	0.002	0.002
Neutral	0.004	0.007	-	0.002	0.002	0.004	-	0.002	0.002	0.002	-	0.005
Happy	0.002	0.002	0.002	-	0.007	0.002	0.002	-	0.004	0.002	0.005	-

Table 3: p-values (2-tail) for the statistical significance test using Wilcoxon signed rank test over the average values of the 3 dimensions obtained for each sample, across emotions. All emotions have statistically different values for the three dimensions as the p-values are less than 0.01.

tive text (speech semantics). The happy emotion is confused with neutral and surprise. There is a lot of confusion when the text emotion and speech emotion have a mismatch, especially with the happy emotion in speech. The happy and sad emotions have a good precision even with poor recall. The Angry emotion has good recall but poor precision, indicating it is confused with others. Some of the descriptive responses relate to happy emotion class, but the forced choice is chosen as neutral. Most of the happy speech with mismatched emotive text is treated as irony.

The results for the three different dimensions are very similar to the results in the earlier section and also have the same correspondence to the human evaluation values distributed with the database. This indicates that the listeners were consistent with their feedback on different dimensions of emotions and the textual emotions did not appear to bias their response. The statistical significance for the emotions across the three dimensions are mentioned in the Table 3 based on the Wilcoxon signed rank test with a significance level of 0.01 for the combined results from Sections 1 and 2 (neutral and emotive text). The table shows that all systems have values that are significantly different from each other. People perceive the emotions as having different values across valence, activation and dominance.

There are three different types of text and speech emotion combinations. When both emotions match, the combination can be termed as a "matched" case. The combinations of Sad / Happy or Angry / Happy can be termed as "opposite" emotion pairs or "strong mismatch", all other combinations can be termed as "ambiguous" (or "mismatch"). The matched, ambiguous and opposite combinations are plotted in Figure 3. The "matched" modality does not lead to a statistically different perception of irony from the "ambiguous" one. However, the "strong mismatch" leads to a significantly stronger sense of irony than the other two conditions (based on Wilcoxon signed rank test, both 2-tail p-values less than 0.01). More precisely, "angry" sentences in a happy voice and "sad" sentences in a happy voice are perceived as the ones with the strongest potential for irony.

This task further emphasizes that listeners are biased by the

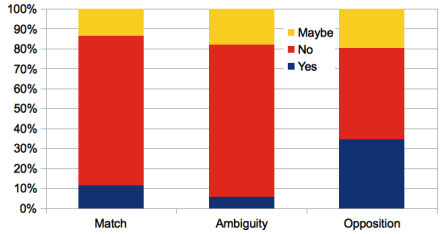

Figure 3: Results for Irony

semantic content of the text for the forced choice selection. The performance is best when the semantic content is matching with the intended emotion in the speech, and irony can be perceived easily when there is a mismatch in speech and text emotions. This supports the hypothesis that listeners tend to be influenced by the semantic content of the text when perceiving emotions. Contradicting this observation, even with opposing emotions in text and speech, some emotions like angry were almost always correctly identified. This also indicates that some emotions such as angry (with very prominent speech factors) are easy to identify even in adverse situations where the semantics in the text do not correspond to the emotion (even when perceiving it as irony).

4.3. Section 3: Context Information

This section provides more details to the listener on the background of the mood of the speaker, through the use of context information spoken in a neutral voice. This information appears to greatly improve the performance, as all emotions appear to be better recognized. As mentioned earlier, one listener did not understand the task and marked all tests as neutral, probably based on the contextual information presented using synthesized speech with a neutral voice. This listener is excluded from the results of this section. As hypothesized the context information has a big influence on the performance. Emotions "sad" and "happy" are significantly better discriminated in section 3, while "angry" is equivalent in both cases. All emotions have good precision with the context information compared to the section 2 without context information. The descriptive response corresponds exactly to the forced choice response. The

Native	German				Non-German			
Emotions	**A**	**S**	**N**	**H**	**A**	**S**	**N**	**H**
Angry	**7**	0	1	0	**35**	0	2	7
Surprise	0	0	0	2	1	0	0	6
Sad	0	**5**	0	0	0	**28**	0	0
Neutral	2	2	**8**	3	0	7	**34**	0
Happy	0	0	0	**4**	0	0	0	**23**
Fear	0	2	0	0	0	1	0	0
Recall	77.8%	55.6%	88.9%	44.4%	97.2%	77.8%	94.4%	63.9%
Precision	87.5%	100%	53.3%	100%	79.5%	100%	82.9%	100%

Table 4: Confusion matrix for forced choice test for Section 4 with listeners with and without German knowledge.

results for different dimensions in this section were also similar to the ones mentioned in the earlier section.

4.4. Section 4: Knowledge of Language

This is an interesting task where people with and without German knowledge were asked to evaluate the emotional speech in German. The text was semantically neutral. It can be seen from the forced choice results in Table 4, that native German speakers perform worse than non-native speakers. Both sad and happy have same precision for native and non-natives with better recall for non-natives. Neutral has better precision and angry has worse precision for non-native German listeners. The number of German speakers is however too small to make any generalizations. The results appear to contradict the hypothesis that the language or culture knowledge improves the emotion identification performance. It is possible the German listeners were biased because of the understanding of the semantic content of the text. The non-native speakers base their judgment only on the signal cues.

The native German speakers described the emotions more closely to the intended emotions, like, “Nervous” or “tired” for sad and “agitated” or “strong” for angry, even though choosing wrong emotion in the forced choice test. The selection of different dimensions were similar to the results in the earlier section.

5. Conclusions

All factors studied in this paper are shown to influence the perception of emotion, with the greatest emphasis on the contextual information and the type of emotion. Happy is usually a very difficult emotion to perceive. The semantics of the text has a great influence on the emotion perception. Though it may not be generalizable from this test, the current results suggests that if the intention is a pure evaluation of emotional content in the speech (comparing specific techniques for emotion generation), it might be better to use listeners without language knowledge to avoid any kind of bias from semantics. If the usability of the emotion in an application is to be checked, it is better to give a full background or context in which the emotional speech appears and listeners may be able to judge better. Forced choice task is the simplest method among the different techniques and does not deviate a lot from the descriptive response case. The happy emotion, which is difficult to perceive, is not discriminated well even in the three dimensions of valence, activation or dominance.

6. Acknowledgments

The work was supported by Eurostars Programme powered by Eureka and the European Community under the project “D-Box: A generic dialog box for multi-lingual conversational applications”.

7. References

[1] Marc Schröder, “Emotional Speech Synthesis: A Review”, Eurospeech, pp. 561-564, 2001.

[2] Heiga Zen and Keichi Tokuda and Alan Black, “Review: Statistical parametric speech synthesis”, Speech Communication, Volume 51(11): pp. 1039-1064, 2009.

[3] Takashi Nose, Junichi Yamagishi, and Takao Kobayashi. “A style control technique for HMM-based expressive speech synthesis”, IEICE Trans. Information and Systems, E90-D(9):1406-1413, 2007.

[4] R. Barra-Chicote, J. Yamagishi, S. King, J. Manuel Monero, and J. Macias-Guarasa, “Analysis of statistical parametric and unit-selection speech synthesis systems applied to emotional speech”, Speech Communication, 52(5):394-404, 2010.

[5] Dimitrios Ververidis and Constantine Kotropoulos, “Emotional speech recognition: Resources, features, and methods”, Speech Communication, Volume 48 (9), pp. 1162-1181, 2006.

[6] Alan W. Black et al., “New Parametrizations for Emotional Speech Synthesis”, Final Report for NPESS team - CSLP John Hopkins Summer Workshop 2011.

[7] Janet E. Cahn, “Generation of Affect in Synthesized Speech”, in Proc. of the Conference of the American Voice I/O Society. Newport Beach, California, 1989.

[8] Cécile Pereira, “Dimensions of emotional meaning in speech”, in Proc. of SpeechEmotion, pp. 25-28, 2000.

[9] Iain R. Murray and John L. Arnott, “Implementation and testing of a system for producing emotion-by-rule in synthetic speech”, Speech Communication, Volume 16 (4): pp. 369-390, 1995.

[10] C. Gobl and A. N. Chasaide, “The role of voice quality in communicating emotion, mood and attitude”, Speech Communication, Volume 40 (1-2), pp. 189 - 212, 2003.

[11] Gregory Hickok and David Poeppel, “Towards a functional neuroanatomy of speech perception”, Trends in cognitive sciences, Volume 4 (4), pp.131 - 138, 2000.

[12] David J. Amante, “The Theory of Ironic Speech Acts” Poetics Today, Volume 2 (2,) Narratology III: Narration and Perspective in Fiction (Winter, 1981), pp. 77-96.

[13] Klaus R. Scherer, “A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology”, in Proc. of ICSLP, China, 2000.

[14] Sungbok Lee et.al., “An Articulatory study of emotional speech production.” Interspeech, Portugal, pp. 497-500, 2005.

[15] Felix Burkhardt et.al., “A Database of German Emotional Speech”, in Proc. of Interspeech, Portugal, 2005.

[16] Roddy Cowie et. al., “FEELTRACE’: An Instrument For Recording Perceived Emotion In Real Time”, in Proc. of ISCA Workshop on Speech and Emotion, pp. 19-24, Ireland, 2000.

[17] M. P. Aylett and C. J. Pidcock, “The cerevoice characterful speech synthesiser sdk”, in AISB, pp. 1748, 2007.

[18] American Heritage, “The American Heritage Dictionary of the English Language”, 4th ed. Houghton Mifflin Company, 2009.

[19] M. P. Aylett and B. Potard and C. J. Pidcock, “Expressive Speech Synthesis: Synthesising Ambiguity”, 8th SSW, Barcelona, 2013.