Emotion Identification for Evaluation of Synthesized Emotional Speech

Stefan Steidl[1]*, Tim Polzehl*[2]*, H. Timothy Bunnell*[3]*, Ying Dou*[4]*, Prasanna Kumar Muthukumar*[5]*,*
Daniel Perry[6]*, Kishore Prahallad*[7]*, Callie Vaughn*[8]*, Alan W. Black*[5]*, and Florian Metze*[5]

[1]Computer Science Department 5, University Erlangen-Nuremberg, Germany
[2] Deutsche Telekom Laboratories / Technische Universität Berlin, Germany
[3] Nemours Biomedical Research, Wilmington, U. S. A.
[4] Center for Speech and Language Processing, Johns Hopkins University, Baltimore, U. S. A.
[5] Language Technologies Institute, Carnegie Mellon University, Pittsburgh, U. S. A.
[6] University of California, Los Angeles, U. S. A.
[7] International Institute of Information Technology, Hyderabad, India
[8] Computer Science Department, Oberlin College, Oberlin, U. S. A.
stefan.steidl@informatik.uni-erlangen.de

Abstract

In this paper, we propose to evaluate the quality of emotional speech synthesis by means of an automatic emotion identification system. We test this approach using five different parametric speech synthesis systems, ranging from plain non-emotional synthesis to full re-synthesis of pre-recorded speech. We compare the results achieved with the automatic system to those of human perception tests. While preliminary, our results indicate that automatic emotion identification can be used to assess the quality of emotional speech synthesis, potentially replacing time consuming and expensive human perception tests.
Index Terms: emotion, speech synthesis, automatic quality assessment, human perception

1. Introduction

In order to improve synthesis of emotional speech, it is necessary to be able to compare different systems and to evaluate their quality. So far, the quality is generally assessed through human perception tests. In order to be able to detect even small differences in the quality of two systems, the number of samples as well as the number of human judges has to be sufficiently high. Finding qualified human participants however is difficult and the number of samples that can be presented to one listener should be limited, to avoid fatigue. Hence, human perception tests are time consuming and expensive. These disadvantages are avoided, if automatic emotion identification could be used as an objective measure to evaluate the quality of emotional speech synthesis. The underlying assumption is that an emotion synthesis system is of high quality, if the intended emotion can be predicted correctly by an emotion identification system that is trained on human voices. Of course, such measures of emotional quality are meant to complement, not replace, existing evaluation metrics such as Mel-Cepstral Distortion (MCD), Mean Opinion Scores (MOS), or others, focusing on naturalness, intelligibility, or accuracy of the synthesized speech.

2. Databases

We used the German "Berlin Database of Emotional Speech" (emoDB) [1] of prototypical emotion portrayals to train and evaluate various emotional speech synthesis systems. 10 actors (5 female and 5 male) produced 10 (emotionally neutral, grammatical, but often non-sensical) sentences each in 7 different emotions: joy (J), neutral (N), boredom (B), sadness (S), disgust (D), fear (F), and anger (A). For our synthesis experiments, we only retained samples which could be identified with an accuracy of at least 80 % in tests with human listeners. Furthermore, the selected samples had to be judged as natural by at least 60 % of the listeners, which leaves 535 samples.

Table 1: Distribution of the seven emotion categories on training and held-out test set (number of samples). Total amount of speech is ca. 25 minutes, which is acceptable for our purposes.

	training	test	$\sum$
joy	63	8	71
neutral	72	7	79
boredom	73	8	81
sadness	52	10	62
disgust	38	8	46
fear	57	12	69
anger	111	16	127
$\sum$	466	69	535

For future experiments, we defined a held-out test set of 69 samples. The classification experiments in this paper are based on the training set of the remaining 466 samples, using the leave-one-speaker-out evaluation method. The distribution of the seven emotion categories for both sets is given in Table 1.

3. Emotion identification system

Our emotion identification system is based on standard state-of-the-art components: we use the openSMILE toolkit [2] for feature extraction and the WEKA data mining toolkit [3] for classification. We focus on easy to extract acoustic features, and use the 1582 acoustic features of the INTERSPEECH 2010 Paralinguistic Challenge baseline [4]. This feature set is obtained by applying a brute-force approach, in which first of all 38 low-level descriptors and their first derivative are computed on the frame level. In a second step, 21 functionals are applied in order to obtain a feature vector of constant length for the whole utterance. Table 2 gives an overview of the low-level descriptors and associated functionals. 16 zero-information features (e. g.

Table 2: Description of the acoustic features based on 38 low-level descriptors and their first derivative and 21 functionals.

Descriptors	Functionals
PCM loudness	position max./min.
MFCC [0-14]	arithm. mean, std. deviation
log Mel freq. band [0-7]	skewness, kurtosis
LSP frequency [0-7]	lin. regression coeff. 1/2
F0 by sub-harmonic sum.	lin. regression error Q/A
F0 envelope	quartile 1/2/3
voicing probability	quartile range 2−1/3−1/3−2
jitter local	percentile 1/99
jitter DDP	percentile range 99−1
shimmer local	up-level time 75/90

the minimum of the fundamental frequency is always zero) are removed from the set of 1596 possible features, and two additional features (*F0 number of onsets* and *turn duration*) are added, resulting in a set of 1582 features.

For classification, we used Support Vector Machines (SVMs) with a linear kernel and Sequential Minimal Optimization (SMO) for learning. The complexity parameter was determined in advance and set to 0.1 for the classification experiments reported in Section 5.2. As the classes are slightly unbalanced, we applied WEKA's implementation of the Synthetic Minority Oversampling Technique (SMOTE). A 10-fold leave-one-speaker-out evaluation was used to determine the performance of the classifier on the whole data set.

For evaluating a synthesized voice sample, the synthesized voice was treated as additional data of the same speaker as some systems are based on the natural parameters (e. g. natural durations) of this speaker. Thus, neither the synthesized voice nor the data of the corresponding human speaker was seen in the training process. Prior to the classification process, a z-score speaker normalization of the features was applied.

Since the number of features (1582) is rather high, we also applied principal component analysis (PCA) and feature ranking based on the information gain ratio (IGR) to reduce the number of features. The results show that the SVM classifier can handle the large number of features and that reducing the number of features does not significantly improve the results.

In order to further analyze the differences between synthesized and human emotional speech, we performed separate classification experiments for different sub-sets of features: we split our feature set into different types using the available low-level descriptors as shown in Table 3. Even though our feature set does not explicitly model word or pause durations, the position of the extreme values of all low-level descriptors are durations, and therefore make up a separate group.

Table 3: Different types of acoustic features and number (count) of low-level features derived.

Type	Number
Prosodic Features	
F0	72
energy	38
durations	154
Voice Quality Features	
jitter	68
shimmer	34
voicing probability	38
Spectral Features	
MFCC	570
MEL	304
LSP	304
	1582

4. Parametric emotional speech synthesis

For comparison of human and machine evaluation, we created five parametric speech synthesis systems with varying degrees of prediction and hence of different quality. We use "ClusterGen" Parametric Synthesis (CGP) [7], as this will use the data more efficiently than any concatenative technique given the amount of type of training data we have. All systems are based on articulatory features as an intermediate representation [6]. Importantly, we have two dimensions on the systems. The "E" systems (*ttsE* and *cgpE*) include explicit emotion information in the training and testing, i. e. the model uses speech labeled as angry to model angry speech. The non-E systems (*tts* and *cgp*) do not use explicit emotion information, thus acting as controls.

The second dimension is changing the amount of information that is predicted, to show the importance of different parts of the signal. The *resynth* system does not predict, but simply decomposes the signal into its components and reconstructs it. *cgpE* and *cgp* use natural durations, and predict spectrum and F0. *ttsE* and *tts* predict F0, spectrum, and durations.

tts Full text-to-speech (TTS) ignoring emotional information in both training and testing. This is a control experiment; the accuracy in the perception and identification experiments should be at chance level (14.3 % for 7 classes).

ttsE As with the *tts* system, this system predicts durations, F0, and spectrum, but also has an "emotion flag" identifying the desired emotion. Training also has this flag, thus the models can generate different emotions. Classification should be better than chance.

cgp As the *tts* system, this system ignores emotions in synthesis and training. It predicts F0 and spectrum, but uses the durations extracted from an original, matching human speech sample. The actual duration patterns are actually dependent on the emotion and – although not modeled explicitly – thus this system actually contains information about the intended emotion of the speaker, and shows the importance of durations.

cgpE As the *cgp* system, this system predicts F0 and spectrum and uses the actual durations from an original speech sample. This system uses emotional labeling in both training and testing, and will generate different predictions for each emotion. We expect the recognition results to be better than the ones for *cgp*.

resynth This system is a pure re-synthesis approach, using natural durations, F0, and spectrum, processed with a speech synthesis framework. It represents an upper limit for the quality of our emotional speech synthesis.

5. Evaluation

In order to evaluate the quality of the automatic assessment, the results of the automatic emotion identification system are compared to the ones obtained in a human perception test.

5.1. Human perception tests

As human perception tests are time consuming and expensive, we selected an emoDB subset that contains 5 randomly selected

Table 4: Results of the human perception tests compared to the results of the emotion identification system for different synthesized voices and the original human voices.

	Human Perception	Emotion ID	
emoDB	subset	subset	full
tts	15.6 %	14.2 %	14.1 %
ttsE	17.5 %	17.1 %	29.0 %
cgp	61.5 %	62.8 %	64.5 %
cgpE	61.0 %	74.2 %	71.5 %
resynth	79.8 %	85.7 %	81.8 %
original	87.7 %	82.8 %	83.7 %

samples for each emotion. For each of the 6 experiments, each of the 35 audio samples was presented to 15 native German listeners in random order using a web interface. For each sample the human judges had to select one of the 7 given emotions, resulting in 525 judgments for each experiment. The judges were mostly students (36% male / 64% female, mean age 26 years, age range 22-39) and wore high-quality headphones, in a quiet office environment. Listening and judging took 9.5 seconds on average per sample.

As expected, the results of the *tts* control experiment (15.6 %) are close to chance level (14.3 %). According to the human judges, there is no significant difference between the *ttsE* and *tts* systems, even though *ttsE* includes emotion information. However, if natural durations are used instead of predicted ones, without an emotion flag (system *cgp*), human listeners are clearly able to distinguish the seven emotions. The accuracy for *cgp* is 61.5 %. Again, adding an emotion flag does not lead to better results in the human perception test, in fact leading to an insignificant degradation for *cgpE* (61.0 % vs. *cgp*'s 61.5 %). In our upper limit experiment – the re-synthesis based on natural durations, F0, and spectrum – the human judges can predict the seven emotions with an accuracy of 79.8 %. This is certainly a good result, but it is still worse than the performance of the human listeners for the original recordings of the actors, which is 87.7 %. The accuracies are summarized in Table 4. The confusion matrices are shown in Table 5.

There appears to be no generalizable systematic effect of the *E* emotion flag. Anger and sadness recognition clearly benefits in the *tts* systems. While fear recognition suffers, all other emotions remain near the baseline. When applied to *cgp* systems, the flag inclusion boosts the recognition of all emotions except sadness. Also, accuracy of neutral speech recognition decreases. Consequently, learning durations, F0 and spectral parameters from emotion-specific data partitions generally improves recognition of synthesized anger. Still, using natural durations gives the biggest improvement for all classes.

5.2. Automatic evaluation based on emotion identification

The emoDB-trained emotion identification system described in Section 3 is now used to evaluate the five systems for synthesis of emotional speech described in Section 4.

For the three systems *tts*, *ttsE*, and *cgp*, the results of the automatic system on the subset are very close to the results of the human perception test. For *cgpE* and *resynth*, better results are obtained with the objective measure, whereas the results are worse for the original human voices. However, this subset of 35 samples is very small and hence the significance of these differences is low. For the performance of the emotion identification system on the whole training set, similar trends can be observed. On the whole training set (which, again, we use for leave-one-speaker-out testing), the system *ttsE* is judged clearly better than *tts*, and *cgpE* is clearly better than *cgp*. *Resynth* clearly represents the best of the five speech synthesis systems,still performing slightly lower than the original human voices, though. Table 6 shows the confusion matrices.

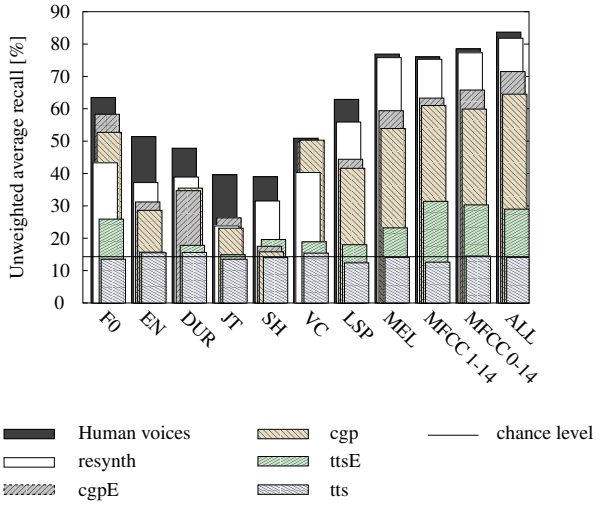

Figure 1: Automatic emotion identification results for different feature types, see Table 2 and Table 3.

Figure 1 shows the performance of emotion identification systems trained on different sub-sets (or types) of acoustic features. In general, the classifiers based on spectral features (MEL, MFCC 0-14, MFCC 1-14) as well as LSP perform very well. They also contain the highest numbers of individual features, which can bias the evaluation. Inclusion of the emotion flag improves synthesis and objective evaluation on basis of these features, as expected. The same can be observed for F0 features. All other features change inconsistently with respect to the switch. The lowest performance is obtained with the small group of jitter and shimmer features. Evaluation based on VC or F0 features only leads to inconclusive results, as our classifiers seem to detect synthesis predictions of pitch and voicing better than actual resynthesized pitch and voicing.

The performance of the *resynth* MEL and MFCC features is almost identical to the performance on human voices. However, a clear drop in performance is observed for F0 features and the voicing probability features of the final F0 candidate. Obviously, there are clear differences between the energy patterns and smaller differences between the durations patterns, too.

6. Conclusions

The results of the emotion identification experiments are very consistent and mostly confirm our intuition. The results of the objective measure highly correlate with the ones of the human perception tests – however at a much lower price, much faster, and with much lower effort involved in the evaluation. It is interesting to note that for both human evaluation and evaluation by automatic classification using natural durations seems to be the most important factor to achieve high accuracy.

Thus, automatic emotion identification can be used successfully to judge the quality of emotional speech synthesis systems, at least for in-development assessment of improvements, if not for final judgments. In addition, the analysis of different feature types can give valuable insights into why synthesis systems perform differently, and worse than human voices.

reference \ hypothesis	J	N	B	S	D	F	A	∑
J	**4**	33	1	8	4	16	9	75
N	4	**46**	5	6	3	7	4	75
B	0	48	**1**	2	3	18	3	75
S	5	46	4	**2**	2	12	4	75
D	1	34	5	9	**4**	17	5	75
F	11	36	0	2	4	**19**	3	75
A	0	41	2	3	7	16	**6**	75
								525

(a) tts: **15.6 %** accuracy

reference \ hypothesis	J	N	B	S	D	F	A
J	**4**	19	2	6	1	42	1
N	15	**28**	3	16	2	9	2
B	1	39	**3**	15	3	12	2
S	4	24	1	**21**	4	19	2
D	1	31	2	16	**5**	18	2
F	8	29	2	13	3	**20**	0
A	3	19	0	6	5	31	**11**

(b) ttsE: **17.5 %** accuracy

reference \ hypothesis	J	N	B	S	D	F	A
J	**40**	18	4	0	1	4	8
N	0	**54**	9	3	0	4	5
B	3	6	**63**	1	1	1	0
S	0	8	16	**47**	1	3	0
D	1	6	3	18	**34**	13	0
F	6	11	0	1	4	**37**	16
A	17	7	1	0	0	2	**48**

(c) cgp: **61.5 %** accuracy

reference \ hypothesis	J	N	B	S	D	F	A	∑
J	**41**	19	0	1	2	4	8	75
N	1	**45**	8	7	3	1	10	75
B	0	6	**67**	2	0	0	0	75
S	0	6	25	**40**	1	3	0	75
D	0	5	1	17	**35**	17	0	75
F	4	15	4	1	0	**44**	7	75
A	21	6	0	0	0	0	**48**	75
								525

(d) cgpE: **61.0 %** accuracy

reference \ hypothesis	J	N	B	S	D	F	A
J	**63**	6	0	0	1	1	4
N	0	**53**	12	7	0	0	3
B	0	0	**68**	5	2	0	0
S	0	1	8	**63**	0	3	0
D	0	2	0	12	**61**	0	0
F	15	0	0	0	0	**46**	14
A	8	2	0	0	0	0	**65**

(e) resynth: **79.8 %** accuracy

reference \ hypothesis	J	N	B	S	D	F	A
J	**69**	1	0	2	1	0	2
N	0	**64**	5	2	1	0	3
B	0	1	**74**	0	0	0	0
S	0	0	5	**69**	0	1	0
D	0	1	2	9	**62**	1	0
F	16	1	0	0	0	**54**	4
A	5	1	0	0	0	0	**69**

(f) original: **87.8 %** accuracy

Table 5: Results of the **human perception tests** for different synthesized voices and the original intended emotions. Using natural durations seems to be important for classification (5 audio files for each of 7 classes, annotated by 15 labelers each = 522 comparisons).

reference \ hypothesis	J	N	B	S	D	F	A	∑
J	**15**	8	12	0	6	16	6	63
N	20	**9**	15	0	3	17	8	72
B	19	9	**13**	0	7	20	5	73
S	13	10	13	**0**	2	14	0	52
D	14	4	5	0	**4**	7	4	38
F	15	8	9	0	5	**15**	5	57
A	31	16	23	0	7	25	**9**	111
								466

(a) tts: **14.1 %** UAR, 13.9 % WAR

reference \ hypothesis	J	N	B	S	D	F	A
J	**33**	5	3	0	0	15	7
N	3	**33**	11	0	19	5	1
B	2	36	**9**	0	18	7	1
S	1	20	12	**0**	16	3	0
D	1	13	4	0	**9**	11	0
F	6	18	3	0	15	**11**	4
A	40	0	0	0	0	16	**55**

(b) ttsE: **29.0 %** UAR, 32.1 % WAR

reference \ hypothesis	J	N	B	S	D	F	A
J	**43**	4	0	0	0	10	6
N	0	**57**	4	0	1	10	0
B	0	18	**43**	2	9	1	0
S	0	3	10	**33**	6	0	0
D	2	4	3	2	**25**	1	1
F	3	7	3	0	1	**41**	2
A	51	1	0	0	4	6	**49**

(c) cgp: **64.5 %** UAR, 62.4 % WAR

reference \ hypothesis	J	N	B	S	D	F	A	∑
J	**47**	3	0	0	0	7	6	63
N	0	**61**	7	0	1	3	0	72
B	0	16	**48**	1	5	3	0	73
S	0	0	10	**39**	2	1	0	52
D	4	3	2	2	**25**	2	0	38
F	2	6	2	0	2	**43**	2	57
A	31	0	0	0	5	9	**66**	111
								466

(d) cgpE: **71.5 %** UAR, 70.6 % WAR

reference \ hypothesis	J	N	B	S	D	F	A
J	**44**	1	0	0	1	8	9
N	0	**62**	5	0	1	4	0
B	0	3	**65**	3	2	0	0
S	0	1	5	**46**	0	0	0
D	0	0	2	0	**32**	2	2
F	6	3	0	1	0	**45**	2
A	23	0	0	0	0	3	**85**

(e) resynth: **81.8 %** UAR, 81.3 % WAR

reference \ hypothesis	J	N	B	S	D	F	A
J	**40**	0	0	0	1	7	15
N	0	**64**	6	0	2	0	0
B	0	4	**66**	2	1	0	0
S	0	0	1	**51**	0	0	0
D	1	4	1	1	**29**	1	1
F	4	2	0	1	1	**47**	2
A	13	1	0	0	0	1	**96**

(f) original: **83.7 %** UAR, 84.3 % WAR

Table 6: Confusion matrices and performance of the **automatic emotion identification system** for different synthesized voices and the original human voices in terms of the (unweighted) average recall (UAR) and the weighted average recall (WAR) / accuracy.

7. Acknowledgments

This paper is based on work which was conducted by the authors during a 2011 Johns Hopkins University Summer Workshop [5]. This workshop was supported by grants from NSF and Google, and made possible by enthusiastic JHU faculty, students, and staff. Additional support was provided by Deutsche Telekom Innovation Laboratories and the German Academic Exchange Service (DAAD).

8. References

[1] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B., "A Database of German Emotional Speech", Proc. INTERSPEECH 2005, Lisbon, Portugal, 2005, pp. 1517-1520

[2] Eyben, F., Wöllmer, M., and Schuller, B., "openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proc. ACM Multimedia, Florence, Italy, 2010.

[3] Witten, I. H. and Frank, E., Data mining: "Practical machine learning tools and techniques", Morgan Kaufmann, San Francisco, 2005.

[4] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., and Narayanan, S., "The INTERSPEECH 2011 Paralinguistic Challenge", Proc. INTERSPEECH 2010, Makuhari, Japan, 2010, pp. 2794-2797.

[5] Workshop "New Parameterization for Emotional Speech Synthesis", Center for Language and Speech Processing (CLSP) at the Johns Hopkins University, Baltimore, `http://www.clsp.jhu.edu/workshops/ws11/groups/npess`, 2011.

[6] Black, A. W., Bunnell, H. T., Dou, Y., Muthukumar, P. K., Metze, F., Perry, D., Polzehl, T., Prahallad, K., Steidl, S., and Vaughn, C.; "Articulatory Features for Expressive Speech Synthesis", Proc. ICASSP 2012, Kyoto, Japan, 2012.

[7] Zen, H., Tokuda, K., and Black, A. W., "Statistical Parametric Speech Synthesis", Speech Communication 51(11), 2009.