Learning with synthesized speech for automatic emotion recognition

Björn Schuller, Felix Burkhardt

Angaben zur Veröffentlichung / Publication details:

Schuller, Björn, and Felix Burkhardt. 2010. "Learning with synthesized speech for automatic emotion recognition." In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 14-19 March 2010, Dallas, TX, USA*, edited by Scott C. Douglas and Nasser Kehtarnavaz, 5150–53. Piscataway, NJ: IEEE.
https://doi.org/10.1109/ICASSP.2010.5495017.

LEARNING WITH SYNTHESIZED SPEECH FOR AUTOMATIC EMOTION RECOGNITION

Björn Schuller[1] *and Felix Burkhardt*[2]

[1]Institute for Human-Machine Communication, Technische Universität München, Germany
[2]Deutsche Telekom Laboratories, Berlin, Germany
schuller@tum.de, felix.burkhardt@telekom.de

ABSTRACT

Data sparseness is an ever dominating problem in automatic emotion recognition. Using artificially generated speech for training or adapting models could potentially ease this: though less natural than human speech, one could synthesize the exact spoken content in different emotional nuances - of many speakers and even in different languages. To investigate chances, the phonemisation components Txt2Pho and openMary are used with Emofilt and Mbrola for emotional speech synthesis. Analysis is realized with our Munich open Emotion and Affect Recognition toolkit. As test set we gently limit to the acted Berlin and eNTERFACE databases for the moment. In the result synthesized speech can indeed be used for the recognition of human emotional speech.

Index Terms— Speech Synthesis, Affective Computing, Emotion Recognition, Speech Analysis

1. INTRODUCTION

If synthesized speech would be suited to train or adapt acoustic models for the recognition of human emotional speech, countless options would open up: not only could the ever-present data sparseness in the field [1] be overcome in general, but emotional speech could be produced for different target groups by age or gender, in different and also sparse resource languages, and fitting to the spoken content at hand. The latter would help overcome the challenge of text-independent emotion recognition: provided a reliable automatic speech recognition one could first recognize the phonetic content, and then re-produce it in various facets for the recognition of emotion. This step can be realized without explicitly producing audio of the synthesized speech. The feasibility was demonstrated successful for a distantly related audio-task: in [2] improved recognition of music chords is shown by synthesis of training material from MIDI symbolic music. Numerous different sound fonts can be used to obtain audio realizations of chords, just as numerous different speakers and texts can be used in our case to produce virtually countless training instances. Whether this concept is generally applicable for the recognition of emotion in speech will be the subject of the present paper: we first generated synthesized speech with matching textual content to the database of human speech by two different phonemisation components (in order to gain as much additional data as possible), namely Txt2Pho and openMary, in combination with Emofilt and Mbrola. Next, we use our open source Emotion and Affect Recognition toolkit [3] to produce a 6 k space of acoustic features. For experiments we decided for our Berlin Emotional Speech Database as introduced in [4] and the eNTERFACE corpus for cross-corpus tests [5]. Extensive test-runs are carried out using Bayesian Networks as classifier of choice.

The structure is as follows: in Sec. 2 and Sec. 3 we provide details on the synthesis and analysis parts prior to the description of the databases in Sec. 4 that serve for the experiments detailed in Sec. 5. The paper ends with our conclusions in Sec. 6.

2. EMOTIONAL SPEECH SYNTHESIS

2.1. Overview

Speech synthesis is usually done in a two step approach. First, the text gets analyzed by a natural language processing (NLP) module and converted into a phonemic representation aligned with a prosodic structure, which is then passed to a digital speech processing (DSP) component in order to generate a speech signal. We developed an emotional speech synthesis system on the basis of Mbrola [6]. In order to obtain as many speech samples as possible, we used two different phonemisation components, namely Text2Pho and openMary [7] for natural language processing. Emofilt acts as a transformer between the phonemisation (Text2Pho or openMary) and the speech-generation component (Mbrola). The emotional simulation is achieved by a set of parametrized rules that describe manipulation of certain acoustic aspects of a speech signal. The rules were motivated by descriptions of emotional speech found in the literature. Before the rules are applied by Emofilt, the input phoneme chain gets syllablised by an algorithm based on sonority hierarchy. In addition, stressed syllables are identified as those that carry local pitch contour maxima [8]. For the experiments we synthesized the 10 sentences of the Berlin Emotional Database (cf. sec. 4), simulated 8 target emotions and emotion-related states (boredom, despair, fear, happiness, hot anger, joy, sadness, and yawning) plus neutral with Emofilt, using all seven German voices for Mbrola (4 female and 3 male), thus getting 1 260 samples ($10 \times 2 \times 9 \times 7$). The following sections describe the modifications provided by EmoFilt.

2.2. Acoustic Parameter Modification Methods

The pitch contour of the whole input as well as selected syllables can be modified by altering either the level, the range or the form of contour.

The speech rate can be modified for the whole phrase, specific sound categories or syllable stress-types separately by changing the duration of the phonemes (given as a percentile). Because with Mbrola the voice quality of the speech is fixed within the diphone inventory, we had to restrict ourselves to jitter (fast fluctuations of the F_0-contour) and vocal effort in terms of voice quality modification. In order to simulate jitter, the F_0 values can be displaced by a percentile alternating down and up. Respecting vocal effort, for the German language exist two voice-databases that were recorded in three voice-qualities: normal, soft, and loud (cf. [9]). Considering articulation

modification, a diphone synthesizer has a very limited set of phoneme realizations and does not provide for a way to do manipulations with respect to the articulatory effort. Thus, the substitution of centralized vowels with their decentralized counterparts and vice versa is possible as a work-around to change the *vowel precision*.

2.3. Simulating Emotional States

The modifications for the eight emotion categories used in the oncoming experiments are, as stated above, inspired by a literature review, manually fine tuned, and partly verified by perception experiments [10]. Emofilt is freely available[1] and the reader is invited to reproduce the simulations.

Of course the emotional expression that is generated by these rules is very prototypical and only one possibility to display the target emotions. In order to get a higher variety it would be possible to randomly shift the parameters for the modifications slightly, or use the Emofilt graded-emotion function which generates stronger or weaker versions of the modification rules.

3. EMOTIONAL SPEECH ANALYSIS

We use a systematic generation of features using our open source feature extraction [3]. In detail, the extended set in comparison to [11] comprises of 39 low-level descriptors as dc offset (DC), extremes (Min, Max), and zero-crossing-rate (ZCR) from the time signal, root mean square (RMS) and logarithmic (LOG) frame energy, pitch (F0, normalised to 500 Hz), strength, and quality as well as harmonics-to-noise ratio (HNR) by autocorrelation function, and Mel-frequency cepstral coefficients (MFCC) 0 – 15. To each of these, the delta and double delta coefficients are additionally computed. Next, 51 functionals such as mean, absolute mean, standard deviation, variance, kurtosis, skewness, minimum and maximum value, relative position, and range as well as two linear and three quadratic regression coefficients with their mean absolute (MAE) and square (MSE) errors are computed per speech turn as given by the databases. Thus, the total feature vector contains $39 \cdot 3 \cdot 51 = 5\,967$ attributes. More details on feature implementation and choice are found in [3], where the exact same set has been evaluated for tests on six standard corpora including the human speech ones investigated herein.

The classifier of choice in this work is a simple Graphical Model in form of a hierarchical Bayesian Network with discrete observation nodes and one top-level parent node which lead to slightly better results than Support-Vector Machines (SVM) as applied in most of our previous investigations (cf. e. g. [11]). For optimal results we found it best to use the Kononenko MDL criterion [12] for feature discretization and not to standardize the features per corpus. Scores are based on entropy as measure throughout model topology optimization. Out of the original large feature space a small subset is found by using correlation-based feature selection with greedy stepwise search leaving the testing corpora out. Thus, between 106 – 157 features are used. These are roughly 42 % cepstral, 29 % spectral, 16 % pitch-related, 7 % directly based on the time signal, 4 % energy related, and the remainder is based on voice quality.

4. EMOTIONAL SPEECH DATABASES

The well known set chosen to test the effectiveness of our emotion classification experiments is the freely available studio recorded Berlin Emotional Speech Database (EMO-DB) [4], which covers *anger*, *boredom*, *disgust*, *fear*, *joy*, *neutral*, and *sadness* speaker emotions. The spoken content is pre-defined by ten German emotionally neutral sentences as *"Der Lappen liegt auf dem Eisschrank" (The cloth is lying on the fridge.)*. Ten (five female) professional actors spoke these repeatedly in each target emotion. We use the 494 speaker turns usually considered in works on this set (e. g. [13]).

For cross-corpus evaluations among human speech sets, we need a secondary set: the eNTERFACE [5] corpus is a further public, yet audiovisual emotion database. It contains induced *anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise* speaker emotions. 42 subjects (eight female) from 14 nations are included. It consists of office environment recordings of pre-defined spoken content in English. Each subject was instructed to listen to six successive short stories, each of them eliciting a particular emotion. Overall, the database consists of 1 170 samples in almost perfect class balance.

5. EXPERIMENTS AND RESULTS

As evaluation measures we employ weighted (WA, i. e. accuracy) and unweighted (UA, thus better reflecting imbalance among classes) average recall as suggested in [11]. We further provide the area under the receiver operating characteristic (ROC). The receiver operating characteristic plots the true positive rate (*TPR*) over the false positive rate (*FPR*) achieved by a binary classifier. In the case of multiple emotions, i. e. classes, this value reflects the average separation ability between one class and the others, i. e. the detection ability. The selection of these measures is justified by the non-equal distribution of instances among classes. The highest possible AUC (area under curve) is 1.0, equal to the whole graph area, and achievable only by a 'perfect' classifier. Random guessing has an AUC of 0.5, since it corresponds to the diagonal line in the ROC space. A reasonable classifier should therefore have an AUC that is significantly greater than 0.5, with better classifiers yielding higher values. All tests are carried out in strict speaker independence. In fact, apart from the first experiment where we investigate whether synthesized speech can outperform speech from the same corpus under ideal conditions, the challenging requirement of complete independence of speakers, room conditions, microphoning, spoken content, and understanding of the target emotions is met by full cross-corpus testing. Overall, we carry out four generally different types of experiments to evaluate the benefit of synthesized speech for acoustic model training and adaptation.

Throughout the first two experiments we use the EMO-DB database as test-set of human emotional speech. We first consider training on synthesized speech versus training on human speech from the same corpus. In the case of human versus human speech speaker independence is preserved by splitting the EMO-DB into a first training partition indexed EMO-DB^1 containing the five speakers with lower subject index, and a respective second partition indexed EMO-DB^2 of the five higher index speakers for testing. Note that cross-validation is avoided for better comparability with the other reported results. In these experiments we consider recognition of six of EMO-DB's original seven emotion classes, as disgust cannot be modelled by the synthesizers at present. Table 1 displays the results. The following mapping of synthesized emotions onto the ones of EMO-DB was chosen: yawning was mapped onto boredom, despair onto fear, and joy onto happiness. The other emotions were mapped straight forward as they are labelled identically. These mappings proved to raise recognition performance in before-hand tests.

Clearly, it was not to be expected that training on synthesized speech could outperform training on human speech from the same corpus, though the same spoken content was synthesized. However,

[1] http://emofilt.sourceforge.net

Table 1. Recognition of six (all but disgust) emotions of the Berlin Emotional Speech Database (EMO-DB): training on human (EMO-DB[1]) or synthesized speech (SYN) and testing on a disjunctive set of speakers from EMO-DB (EMO-DB[2]).

Train	Recall [%]		Prec. [%]		AUC	
	UA	WA	UA	WA	UA	WA
6-class						
EMO-DB[1]	76.1	79.2	83.6	81.1	0.95	0.95
SYNTH	62.0	58.4	66.1	68.1	0.91	0.90

Table 2. Recognition of 4 (anger, fear, sadness, happiness), and 3 (excluding fear) emotions of the Berlin Emotional Speech Database (EMO-DB): training on human (eNTERFACE) or synthesized speech (SYN).

Train	Recall [%]		Prec. [%]		AUC	
	UA	WA	UA	WA	UA	WA
4-class test on EMO-DB						
eNTERFACE	38.8	47.2	53.2	54.7	0.81	0.80
SYNTH	54.9	61.9	74.0	68.7	0.86	0.84
eNTER+SYNTH	55.3	64.2	60.1	59.9	0.86	0.86
3-class test on EMO-DB						
eNTERFACE	62.6	64.8	70.2	69.1	0.83	0.81
SYNTH	72.6	75.4	75.7	73.5	0.89	0.88
eNTER+SYNTH	75.0	79.5	81.7	79.3	0.91	0.90

the recall and precision rates seem promisingly high. We thus next experimented with combination of human and synthesized speech in order to see whether we can gain accuracy by increase of the amount of training material. Sadly, however, no gain could be obtained: 1:1 inclusion of human and synthesized speech resulted in a downgrade over using only human speech. Gradually increasing the up-sampling by a repetition factor up to as high as 24:1 in terms of human:synthesized speech we observed an asymptotic approach to training with only human speech. However, this could never be surpassed.

We thus next considered a more fair comparison: in fact we carried out cross-corpus evaluation when training on synthesized and testing on human speech. This is a only sparsely researched and difficult task. In our second experiment we therefore take the eNTERFACE corpus into play preserving the previously chosen mappings: from now on we trained on either eNTERFACE or synthesized speech or their combination and test exclusively on the EMO-DB with all speakers. However, the set of emotions was reduced to 4 classes, as eNTERFACE does not contain neutral or boredom as classes. We next excluded "fear" which is apparently badly modelled both by eNTERFACE and the synthesized speech. Table 2 depicts all respective results.

We next consider testing on eNTERFACE and training on either EMO-DB or synthesized speech. This has the interesting consequence that now the human speech and the synthesized speech are of the same language and contain the same spoken content. They both have to deal with the target set being in a different language and thereby naturally of different phonetic content. Table 3 shows the respective results as before. Note that now unweighted and weighted measures tend to be close to each other, as eNTERFACE is almost fully balanced in terms of classes. Further note that in this experiment the choice of classifier was observed to be an influential factor: SVM delivered slightly

Table 3. Recognition of 4 and 3 emotions (cf. Table 2) of the eNTERFACE database: training on human (EMO-DB) or synthesized speech (SYN).

Train	Recall [%]		Prec. [%]		AUC	
	UA	WA	UA	WA	UA	WA
4-class test on eNTERFACE						
EMO-DB	40.3	39.8	46.6	46.8	0.72	0.72
SYNTH	49.0	49.0	51.8	51.5	0.75	0.75
EMO-DB+SYNTH	51.6	51.6	51.9	51.9	0.76	0.76
3-class test on eNTERFACE						
EMO-DB	54.4	54.0	61.1	61.1	0.79	0.79
SYNTH	58.3	58.0	57.7	57.6	0.80	0.79
EMO-DB+SYNTH	61.9	61.5	60.8	60.8	0.81	0.81

better results (not reported) when training on human speech for this exact experiment while the synthetically produced speech used here seems to better be modelled when statistical classifiers (our Bayesian Network) are used. Clearly, this depends on the implementation of the parameter variation which seems to be well related to Gaussian modelling. While this impacted the recall rates, the area under curve was practically unaffected. Also, this effect was only observed for training on EMO-DB and testing on eNTERFACE and not vice versa.

So far, both phonemisation components were used together. However, one could use the training on synthesized speech and testing on human speech as potential guideline for the quality of the synthesis result. We therefore next provide recognition results for each component individually on the two sets of human speech. EMO-DB thereby served for test with matched spoken content and eNTERFACE for a phonetically independent setting, respectively. For better comparability between these two different settings, we limit the number of target emotions to the 4 emotions contained in both considered target sets of human speech. Results for the 4- and 3-class tasks as before are found in Table 4. Finally, we inversed test and train in principle, and observed how training on human speech will react when confronted with synthesized speech. This resembles an 'out-of-the-box' measurement of synthesized speech quality with respect to emotion without the need of training with speech of the synthesizer. Interestingly, anger is hardly recognized here, while before (i. e. when training on synthesized speech and testing on human speech), fear had been troublesome. We thus do not consider the fear reduced set in the results portrayed in Table 5.

6. CONCLUSION

We considered a completely novel approach to the generation of ever-sparse learning material for emotion and affect recognition: using synthesized speech for training and adapting acoustic models. To summarise, the first step was made with surprising success: for cross-corpus tests on acted speech high benefit could be proven, as we will detail out in the following. Significance refers to a one-tailed test performed on unweighted average recall. When training and testing from the same human speech corpus, no benefit in adding synthesized speech could be found. Also, training exclusively on synthesized speech clearly fell behind in direct comparison. This seems to be expected, as the synthesized speech does not take room acoustics, noises present, and microphone characteristics into account. Also, potentially deviating understanding of the emotions may exist. We thus next shifted to cross-corpora analyses which are considerably closer to a real-life application of an emotion recognition system:

Table 4. Recognition of 4 and 3 emotions (cf. Table 2): training on synthesized speech using openMary or Txt2Pho for phonemisation.

Train	**Recall [%]**		**Prec. [%]**		**AUC**	
	UA	WA	UA	WA	UA	WA
4-class test on EMO-DB						
MARY	56.0	63.2	75.2	69.7	0.87	0.86
TXT2PHO	57.2	63.2	50.0	52.1	0.84	0.82
3-class test on EMO-DB						
MARY	71.6	75.4	76.0	73.4	0.90	0.89
TXT2PHO	77.1	77.5	78.4	77.2	0.89	0.87
4-class test on eNTERFACE						
MARY	46.0	45.9	48.6	48.4	0.73	0.74
TXT2PHO	49.8	49.9	53.0	52.7	0.75	0.75
3-class test on eNTERFACE						
MARY	55.1	54.7	55.4	55.3	0.79	0.79
TXT2PHO	60.8	60.5	60.4	60.4	0.80	0.80

Table 5. Recognition of 4 and emotions (cf. Table 2): testing on synthesized speech using openMary or Txt2Pho for phonemisation.

Test	**Recall [%]**		**Prec. [%]**		**AUC**	
	UA	WA	UA	WA	UA	WA
4-class train on EMO-DB						
MARY	54.4	56.8	68.2	66.0	0.84	0.82
TXT2PHO	43.1	40.8	37.2	40.8	0.75	0.71
4-class train on eNTERFACE						
MARY	48.5	55.4	62.5	59.2	0.82	0.79
TXT2PHO	48.8	54.3	59.2	56.5	0.80	0.78

independent of speakers, acoustic, coding, and transmission influences emotions should be assigned according to a more 'general' understanding of emotions, at least as long as they are rather prototypical. In the cross-corpus experiments we found a true surprise: synthesized speech was found to be the better choice, not only if the spoken content was matched (significance levels of 0.001 for the 4-class and 0.01 for the 3-class task), but even if synthesized speech and human in training were both containing the same, yet different phonetic content to testing (significance levels of 0.001 for the 4-class and 0.1 for the 3-class task). Adding human speech to the synthesized speech generally further improved the recall and precision rates and raised the significance level over only human speech; however, no statistical significance was found over synthesized speech without human speech. Thus, in our tests synthesized speech proved the optimal choice throughout the cross-corpora experiments. While differences among the measurements may depend on the chosen classifier, the tendency was not disrupted in our experiments. We next investigated differences between the two phonemisations used. Here, apart from the 4-class task on EMO-DB, significantly better results were obtained using Txt2Pho (level 0.1 to 0.05). In fact, using only one phonemisation would have been the better choice if one compares Tables 4 and 5 with the lines labelled 'SYNTH' in Tables 2 and 3, where both where used, which however lead to a downgrade. Thus, mixing of different phonemisation was not found beneficial in our experiment, and the above described trends are even amplified if the better phonemisation component would have been used, exclusively. As the Tables 4 and 5 show, no significant difference in terms of unweighted average recall could be found whether one trains on synthesized speech and tests on human speech or vice versa. Only one exemption is observed: the Txt2Pho phonemisation falls significantly behind openMary when the training is carried out on EMO-DB. However, clear differences among the emotions were observed. In terms of AUC and accuracy however, better rates are seen when training on human speech and testing on synthesised speech.

For a first proof of concept, we had limited our data choice to acted emotion. Naturally, manifold subsequent steps are left for future research prior to drawing general conclusions: fore mostly, tests with spontaneous speech and naturalistic emotions of low prototypicality.

7. REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[2] K. Lee and M. Slaney, "Automatic chord recognition from audio using a supervised HMM trained with audio-from-symbolic data," in *Proc. 1st ACM workshop on Audio and music computing multimedia table of contents*, Santa Barbara, CA, USA, 2006, pp. 11 – 20, ACM.

[3] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR – Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit," in *Proc. Affective Computing and Intelligent Interaction (ACII)*, Amsterdam, The Netherlands, 2009, IEEE.

[4] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 1517–1520, ISCA.

[5] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," *IEEE Workshop on Multimedia Database Management*, 2006.

[6] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. Van der Vreken, "The mbrola project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes," in *Proc. ICSLP*, 1996.

[7] M. Schröder and J. Trouvain, "The german text-to-speech synthesis system mary: A tool for research, development and teaching.," *International Journal of Speech Technology*, pp. 365–377, 2003.

[8] F. Burkhardt, "Emofilt: The simulation of emotional speech by prosody transformation," in *Proc. Interspeech 2005*, Lisbon, Portugal, 2005.

[9] M. Schröder and M. Grice, "Expressing vocal effort in concatenative synthesis," in *Proc. 15th International Conference of Phonetic Sciences*, Barcelona, Spain, 2003, pp. 2589–2592.

[10] F. Burkhardt and W. F. Sendlmeier, "Verification of acoustical correlates of emotional speech using formant-synthesis," in *Proc. of the ISCA Workshop on Speech and Emotion*, 2000.

[11] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," in *Proc. Interspeech*, Brighton, UK, 2009, ISCA.

[12] I. H. Witten and E. Frank, *Data mining: Practical machine learning tools and techniques, 2nd Edition*, Morgan Kaufmann, San Francisco, 2005.

[13] H. Meng, J. Pittermann, A. Pittermann, and W. Minker, "Combined speech-emotion recognition for spoken human-computer interfaces," in *Proc. International Conference on Signal Processing and Communications*, Dubai, United Emirates, 2007, IEEE.