Augmenting Affect from Speech with Generative Music

Gerhard Johann Hagerer
Institute for Informatics
Technische Universität München
gerhard.hagerer@tum.de

Stefan Ehrlich
Institute for Cognitive Systems
Technische Universität München
stefan.ehrlich@tum.de

Michael Lux
Institute for Informatics
Technische Universität München
michael.lux@tum.de

Prof. Gordon Cheng
Institute for Cognitive Systems
Technische Universität München
gordon@tum.de

CHI'15 Extended Abstracts, Apr 18–23, 2015, Seoul, Republic of Korea.
ACM 978-1-4503-3146-3/15/04.
http://dx.doi.org/10.1145/2702613.2732792

Abstract
In this work we propose a prototype to improve interpersonal communication of emotions. Therefore music is generated with the same affect as when humans talk on the fly. Emotions in speech are detected and conveyed to music according to music psychological rules. Existing evaluated modules from affective generative music and speech emotion detection, use cases, emotional models and projected evaluations are discussed.

Author Keywords
Affective Computing, Emotion Recognition, Speech Analysis, Generative Music, Circumplex Model

ACM Classification Keywords
H.5.m [Information interfaces and presentation (e.g., HCI)]: Miscellaneous.

Introduction
When humans talk with each other several kinds of information are transmitted from one person to the other. Modern communication models assume interpersonal communication is far more than the plain meaning of words being said. Additional information regarding relationship and emotional state of persons puts a message into its right context making its meaning clear[30].

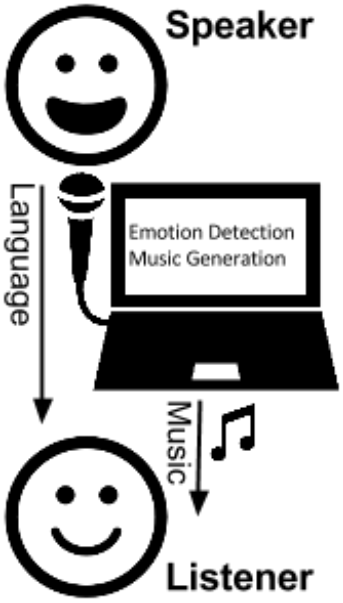

Figure 1: Illustration of the prototype concept. Listener listens in parallel to speaker and to augmented music delivering a congruent affect as the speakers's voice.

One way to express this is by prosodic and spectral features of our voice, which stand out due to their similarity to music and its relation to emotional content. In that regard recent works show evidence that both for vocal and musical expression familiar emotional cues are used. Thus, musical instruments are perceived by humans as "superexpressive voices"[12]. From these results it can be concluded that at least several basic emotions can be expressed and accordingly perceived in two different auditory media: speech and music.

The question arises if the latter can support or even replace the former, i.e. if *emotions from speech can be communicated by corresponding musical accompaniment* as depicted in Figure 1. In fact, there are persons being impaired to decode emotions out of prosody occurring during interpersonal communication while at the same time being capable of understanding the semantic content from speech. This is the case for people affected by receptive aprosodia[20] and indications exist e.g. for schizophrenia[16], autism spectrum disorder[10, 23, 17] and Borderline personality disorder[18]. Since music additionally affects deeper and eventually less damaged or problematic brain regions than speech features[20, 13], those impairments may be surmountable by choosing a "different channel" to deliver the "same code"[12].

This appears useful for non-impaired persons, too, since the qualitative value of spoken audible content raises with its intelligibility. If it can be increased by the suggested method, the door is open for a whole range of new use cases, where emotion in speech is a key element. Spoken audible media as well as self-awareness and communication support in psychology could furthermore benefit from additional levels of sensible meaning for more emotional insight[9, 27].

It is the aim of our work to show a first proof of concept in that regard by proposing a software prototype recognizing affect out of speech and converting it online into emotionally congruent generative music. The aim of this concept is to produce either a musical equivalent or an extension of the emotional information inherent to prosodic and spectral features from speech. The music thereby shall be produced during talking as soon as emotions are recognizable.

Despite actual successes regarding emotion recognition from speech and affective music generation, these two fields have – to the authors' best knowledge – not been combined yet, which is the major contribution of this work. Furthermore, a scientific context around this idea is created with a compilation of related research domains. This involves discussion about an emotion processing pipeline from speech to music, incorporated psychological models and solutions for latencies and misclassifications. Knowing the whole complexity of human emotionality being hardly to abstract by computer models, our assumptions rely on sufficiently evaluated methods breaking it down to a level within which emotions can be handled reliably, reproducibly and consistently.

Related Work

Emotion recognition is the first step, which in terms of speech analysis is still a challenging research problem. Successful methods in that regard are based on machine learning techniques mapping extracted features from audio data to several discrete emotion labels. Research thereby is concerned with questions about which features are most relevant for emotion classification and which classification algorithms yield best results for different use cases. Therefore various studies have already been conducted concentrating on specific problem domains[19, 2], whereby

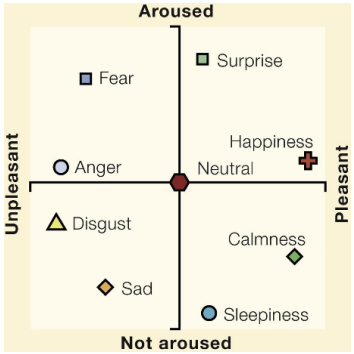

Figure 2: Basic emotions within the Circumplex Model[3, 21]

multiple feature configurations have been prepared to capture emotional information in conjunction with single and hybrid classifiers using openSMILE[7] tool. It takes account of source, spectral, prosodic and voice quality features, which are representative of the vocal tract system and glottal activity during speech and are used extensively in speech analysis tasks. This tool is a result of the INTERSPEECH 2009[24] & 2010[25] challenges, where a baseline for relevant features required for emotion and affect recognition was set. Regarding classification, successes were achieved among others by using Gaussian Mixture Models yielding emotion detection accuracies from 40% to 60% dependent from the respective emotion[14]. Improved classifiers are incorporated in our prototype – see chapter about Methodology.

The mentioned classifiers represent emotions by returning likelihood values for each available emotion label, whereby the most likely appears as the best choice. Another way of representation is the Circumplex Model of Russel[21, 22], which stands out due to its significance for the perception of music – see Figure 2 as example. It suggests a 2D coordinate system, where valence (x-axis) means if an emotion is experienced as pleasant (positive) or unpleasant (negative). Arousal (y-axis) stands for the degree of involved stress and movement. This model was introduced into musicology by Gabrielsson & Lindström[8] and also used by Gomez et al.[11], who showed functional relationships between experienced emotion and structural and expressive features in music.

These findings were then adopted from computer scientists to automatize the targeted generation and manipulation of music pieces with respect to affect. Wallis for example demonstrated an algorithm, which generates music that was assessed by subjects equally as the intended emotion respectively its inherent valence arousal input parameters that were passed to the algorithm[28, 29]. Similar could be achieved by manipulating score and performance features of existing music pieces so, that the expressed emotion changed in an intended way[15]. In view of that, software can generate music with previously determined affect.

Methodology

Emotion Recognition

For emotion recognition based on real time analysis of speech, we firstly describe an existing prototypical classifier trained offline on datasets of emotionally spoken and correspondingly labeled audio data[5, 6]. Later on, an extension for online use will be layed out.

To detect emotions out of prosodic and spectral features, a prototypical classifier will be used designed by Schulze[26] and implemented by Baghel[4]. Its original aim was to run on mobile devices and to detect interruptibility of users during conversations by analyzing their emotions occurring in speech. They improved the recognition rate compared to existing classifiers by extending feature sets defined in INTERSPEECH 2009 & 2010 with new features like formants, their position and bandwidth. Based on these, accuracy rates of 78.64% on Berlin-EMO dataset and 63.39% on FAU-Aibo dataset on average were reached. The results refer to audio signals without background noise, as the aim is to show a proof of concept in ideal situations like quiet environments.

For prototypical online enhancement chunks (windows) of a fixed length and sampled by a fixed update rate are taken from a audio stream like from a microphone. These chunks can be analyzed and classified on the fly yielding frequent emotion labels in time according to the update

Figure 3: Processing pipeline for speech analysis and emotion recognition when a person speaks; continued on next page.

rate. It is not considered as technical problem due to detection times up to 250ms with update rates of 20Hz and large window sizes on actual desktop hardware.

Choosing a correct window size is important. If it is too small, important feature details of an emotional utterance will get lost. If it is too big, the probability of more than one emotion occurring within one window is too high leading to worse accuracy and reactivity. A good compromise can be found by pseudo online testing. Therefore training data is played back in one sequence and processed by the online enhanced classifier. By doing so, misclassifications and delays in system reactivity can be improved by adapting the window size.

First experiments on training data grouped by emotion labels showed accuracies 10% lower than in normal offline testing. Thereby misclassifications are likely to occur at the end of emotion changes due to other audio data filling up the window like pauses or consecutive utterances with different expressivity. Concepts to avoid these are omitting windows with pauses or multiple speakers by utilizing speaker recognition. Otherwise, flickering of the recognized emotional state can be smoothed out so, that sudden but persistent changes as well as slow transitions of emotions in speech are preserved by taking advantage of the continuous character of the 2D Circumplex Model.

A reaction time between occurrence of emotion in speech and generated music is inherent to the prototype due to window size, processing time, smoothing and reaction time of the music generator. We aim at reduction down to 1s. Additionally, extraction of quick detectable speech features like volume are used to change generated music directly by bypassing the emotion processing pipeline, to increase the perceived reactivity. Evaluations take into account how far this is experienced as problematic by subjects – see Evaluation.

Model Transformation

It is necessary to find a mapping between the emotion representation of speech classifiers and the Circumplex Model, which robust music generators like the one used for this work are based on – see Related Work. Therefore Russel deduced unique locations of many emotion labels in Circumplex Space as exemplified in Figure 2.

If related likelihoods from actual speech classification lie on hand, they can be used as probabilistic weights to average all Euclidean 2D positions of the emotion labels. This operation yields a representation of the recognized affect from speech within the Circumplex Model as two dimensional valence arousal position and can be passed to described affective music generation algorithms.

Music Generation

When recognized emotions from speech are represented by 2D positions in the Circumplex Model, these valence arousal values can be passed to the incorporated music algorithm conveying them to accordingly perceived music. Detailed transformational rules therefore are layed out in previously noted research like from Gomez as mentioned in Related Work. Thereby a Midi pattern will be generated by a state machine, whose state transitions depend on settings of several music structural parameters onto which valence and arousal is mapped. Then the Midi patterns are translated into sound by virtual instruments from external software, for example ProTools.

Prototype Evaluation

As denoted, the aim is to show a proof of concept for supporting expressivity of speech. Success hereby is affect from speech and generated music being perceived as

Figure 4: Processing pipeline for transforming recognized emotion to music[1].

equal. If that condition is met, emotional information from voice can alternatively be transmitted by the proposed approach. This is necessary before further inquiries appear reasonable about if this form of augmentation is regarded as helpful by listeners to understand emotions of talking persons.

For a first evaluation, music is generated based on speech audio files from training data. Therefore spoken samples, that are labeled with the same emotion, are played back in a sequence to which music is created and recorded. Later on, these recordings are presented to subjects assessing their inherent affect via questionnaires. By that data, perceived affect from generated music is compared with given emotion labels from training data. The music should be assessed according to related labels.

In a second step, speech data is prepared and music is generated the same way as in previous evaluation. In contrast to before, music recordings and training sample sequences are superimposed and played together at once. Thereby music and sample sequences are combined arbitrarily. Subjects then estimate to what extent music and vocally expressed emotion are congruent and if the combination is meaningful. If that is statistically rather the case for music and speech data belonging together, their technical connection can be stated as perceivable.

The ultimate evaluation goal is a real time online experience, where affective speech unknown to the classifier and music generated to that are presented at the same time. Actors therefore stage emotional content with special respect to vocal expressivity processed by our prototype to create music instantly. The actor himself afterwards rates if the generated music was congruent to his expressed emotion. The same is done by subjects listening to recordings of the musically supported stage.

Acknowledgements

We thank Jitin Kumar Baghel and Florian Schulze from AICOS chair at Technische Universität München for providing us emotion recognition software and advice.

References

[1] Icons by SimpleIcon.com&Flaticon.com, CC BY 3.0.
[2] Anagnostopoulos, C. N., Iliou, T., Giannoukos, I. Features and classifiers for emotion recognition from speech: a survey from 2000–2011. Springer, 2012.
[3] Anderson, D. J., Adolphs, R. A framework for studying emotions across species. *Cell 157*, 1 (2014), 187.
[4] Baghel, J. K. Audio-based characterization of conversations. Master's thesis, Technische Universität München, Institut für Informatik.
[5] Bartels, A., Rolfes, M., Burkhardt, F., Technical University Berlin. Berlin database of emotional speech, requested on 20/11/2014 11:30am.
[6] Batliner, A., Steidl, S., Nöth, E. Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus. In *Proc. of a satellite workshop of LREC* (2008), 28–31.
[7] Eyben, F., Weninger, F., Groš, F., Schuller, B. Recent developments in opensmile, the munich open-source multimedia feature extractor. 835–838.
[8] Gabrielsson, A., Lindström, E. The influence of musical structure on emotional expression.
[9] Galizio, M., Hendrick, C. Effect of musical accompaniment on attitude: The guitar as a prop for persuasion. *Journal of Applied Social Psychology 2*, 4 (1972), 350–359.
[10] Ghaziuddin, M. Defining the behavioral phenotype of asperger syndrome. *Journal of Autism and Developmental Disorders 38*, 1 (2008), 138–142.

[11] Gomez, P., Danuser, B. Relationships between musical structure and psychophysiological measures of emotion. *Emotion 7*, 2 (2007), 377.

[12] Juslin, P. N., Laukka, P. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin 129*, 5 (2003), 770.

[13] Koelsch, S. Brain correlates of music-evoked emotions. *Nature Reviews Neuroscience 15*, 3 (2014), 170–180.

[14] Kostoulas, T., Ganchev, T., Lazaridis, A., and Fakotakis, N. Enhancing emotion recognition from speech through feature selection. In *Text, speech and dialogue*, Springer (2010), 338–344.

[15] Livingstone, S. R., Muhlberger, R., Brown, A. R., Thompson, W. F. Changing musical emotion: A computational rule system for modifying score and performance. *Computer MusicJournal 34* (2010), 41.

[16] Marjolijn H., Ren S., Pijnenborg, K. & M., Aleman, A. Impaired recognition and expression of emotional prosody in schizophrenia: Review and meta-analysis. *Schizophrenia Research 96*, 13 (2007), 135 – 145.

[17] McCann, J., Peppe, S. Prosody in autism spectrum disorders: a critical review. *International Journal of Language & Communication Disorders 38*, 4 (2003), 325–350.

[18] Minzenberg, M. J., Poole, J. H., Vinogradov, S. Social-emotion recognition in borderline personality disorder. *Comprehensive Psychiatry 47*, 6 (2006), 468.

[19] Rao, K. S., S. G. Koolagudi. *Robust Emotion Recognition using Spectral and Prosodic Features*. Springer, 2013.

[20] Ross, E. D. Affective prosody and the aprosodias.

[21] Russell, J. A. A circumplex model of affect. *Journal of personality and social psychology 39*, 6 (1980), 1161.

[22] Russell, J. A., Weiss, A., Mendelsohn, G. A. Affect grid: A single-item scale of pleasure and arousal.

[23] Saulnier, C. A., Klin, A. Brief report: social and communication abilities and disabilities in higher functioning individuals with autism and asperger syndrome. *Journal of autism and developmental disorders 37*, 4 (2007), 788–793.

[24] Schuller, B., Steidl, S., Batliner, A. The interspeech 2009 emotion challenge. *INTERSPEECH*, 312–315.

[25] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Muller, C., Narayanan, S. S. The interspeech 2010 paralinguistic challenge. In *In Proceedings of InterSpeech* (2010).

[26] Schulze, F., Groh, G. Studying how character of conversation affects personal receptivity to mobile notifications. In *CHI'14 Extended Abstracts*, ACM (2014), 1729–1734.

[27] Stratton, V. N., Zalanowski, A. H. Affective impact of music vs. lyrics. *Empirical Studies of the Arts 12*, 2 (1994), 173–184.

[28] Wallis, I., Ingalls, T., Campana, E. Computer-generating emotional music: The design of an affective music algorithm. *DAFx-08, Espoo, Finland* (2008), 7–12.

[29] Wallis, I., Ingalls, T., Campana, E., Goodman, J. A rule-based generative music system controlled by desired valence and arousal. In *Proceedings of 8th international sound and music computing conference (SMC)* (2011).

[30] Watzlawick, P., Bavelas, J. B., Jackson, D. D. *Pragmatics of human communication: A study of interactional patterns, pathologies and paradoxes*. WW Norton & Company, 2011.