Augmenting Emotions from Speech with Generative Music:

Emotion Transformation & Prototype Evaluation

IDP Report

by BSc Gerhard Hagerer

born on November 10th 1986
living at:
Buschingstraße 7
81677 Munich
Germany
Tel.: 015154729954

Chair for
COGNITIVE SYSTEMS
Technische Universität München

Prof. Dr. Gordon Cheng

Advisor:	M.Sc. Stefan Ehrlich
Start:	01.10.2014
Delivery:	15.09.2015

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Location, Date *Signature*

Abstract

The present essay is an activity report about an interdisciplinary student project concerned with a software prototype recognizing affect from human speech and transforming it to congruently perceived generative music. Besides motivating relevant use cases and research from domains like medical sciences and human computer interaction, an emotion recognition and processing pipeline is proposed to map recognized emotions into the Circumplex Model as a pre-stage for music generation. For the latter, existing music algorithms and related research are outlined and an object oriented implementation demonstrated, which is capable of composing and playing music dynamically according to according emotional information. The complete prototype, including emotion recognition from speech plus music generation and all steps in between, was evaluated in a user study. Its results are outlined at the end of this work. They strongly indicate that music created by the utilized approach is perceived as emotionally similar to affective speech. Concludingly, further necessary work is discussed to make the prototype ready for use in real world scenarios, that are especially related to research in medical and psychological domains as well as solutions from media industry for supportive musical accompaniment of artistical verbal stages.

Zusammenfassung

Der vorliegende Aufsatz ist ein Tätigkeitsbericht über ein interdisziplinäres studentisches Projekt über einen Software-Prototyp, der Affekt aus menschlicher Sprache erkennt und dazu passende Musik generativ erzeugt. Dafür werden relevante mögliche Anwendungsfälle in der Medizin und Mensch-Computer-Interaktion motiviert. Es wird ein Verarbeitungsmodell skizziert zur Erkennung von Emotion aus Sprache und dessen Umwandlung hin zu psychologischen Parametern (Circumplex Model) für die Musikerzeugung. Bezüglich letzterem werden existierende Musikalgorithmen und damit verwandte Forschung aufgezeigt und eine eigene objektorientierte Implementierung demonstiert, welche in der Lage ist Musik dynamisch in Entsprechung zu angegebener Emotionsinformation zu komponieren und abzuspielen. Der vollständige Prototyp, einschließlich Emotionserkennung aus Sprache plus Musikgeneration und alle Schritte dazwischen, wurde in einer Nutzerstudie evaluiert. Deren Resultate werden in dieser Arbeit am Ende zusammengefasst. Daraus ergeben sich starke Indizien dafür, dass die Musik des verwendeten Prototyps emotional als ähnlich zur begleiteten Sprache wahrgenommen wird. Abschließend wird erläutert, welcher zusätzlicher Arbeit es noch bedarf um den Prototyp in realen Situationen zu verwenden, insbesondere in der psychologischen bzw. medizinischen Forschung als auch als Endprodukt der Medienindustrie für unterstützende musikalische Begleitung von verbalen künstlerischen Auftritten.

Contents

Preface

This document is the report about an Interdisciplinary Project (abbr. IDP) conducted by Michael Lux and me, Gerhard Hagerer, during the period of October 2014 until September 2015 in the context of our Computer Science studies at Technische Universität München. There are several ideas and techniques, which are not our own but used by us. Firstly, our ideas about emotion recognition from speech were based upon the Master Thesis of Jitin Kumar Baghel and his advisor MSc Florian Schulze from the Informatics Department at TUM [1]. Furthermore, the basis for our generative music algorithm comes from MSc Stefan Ehrlich [2], who has put much effort in kindly and patiently advising us for the success of this project, especially regarding our publication at ACM CHI 2015 [3]. Therefore, we want to explicitly thank all mentioned persons for sharing their knowledge and time with us.
Our IDP consists out of two parts. One part was implemented by Michael Lux and is concerned specifically with details about emotion recognition from speech signals and computationally mapping emotions to two-dimensional coordinates. Thus this part is not covered by the present work. Instead, the overall prototype idea is outlined and motivated as it was already done in a similar way [3]. Additionally a new music generation algorithm and an evaluation of the complete prototype is described here. Splitting the IDP in these two parts was done in accordance with the application documents of the IDP.

Chapter 1

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 to be 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 [4].
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" [5]. 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 1.1. 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 [6] and indications exist e.g. for schizophrenia [7], autism spectrum disorder [8, 9, 10] and Borderline personality disorder [11]. Since music additionally affects deeper and eventually less damaged or problematic brain regions than speech features [6, 12], those impairments may be surmountable by choosing a "different channel" to deliver the "same code" [5].
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 supporting emotions from speech with according music, the door is open for a whole range of new use cases, where intelligible communication of 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 [13, 14].
During our IDP we could implement such a software prototype recognizing affect

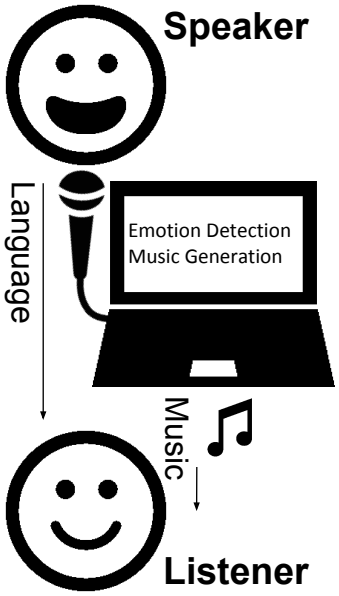

Figure 1.1: A short depiction of what the implemented prototype does: While a person is speaking, the computer is listening and generating music in accordance to the speaker's emotion.

out of speech and converting it online into emotionally congruent generative music. The aim was to produce either a musical equivalent or an extension of the emotional information inherent to prosodic and spectral features from speech. The music is produced dynamically during talking as soon as emotions are recognizable.
An overview of the overall prototype and the data flow within the implementation is given in Chapter 3. As emotion recognition is conducted by Michael Lux, this processing step is omitted. The theoretical basis about how emotion labels from emotion recognition are mapped to two-dimensional numerical values is lined out in Chapter 4. Theoretical and practical details about the implemented generative music algorithm are given in Chapter 5. An evaluation described in Chapter 6 shows, that the stated prototype depicted in Figure 1.1 creates music being perceived as congruent to the emotion from speech.

Chapter 2

Related Work

To further motivate the work done in the present IDP project, the current state of research in the field of augmentive affective computing will be presented in this chapter. Therefore it shall be noted beforehand, that there are various ways of recognizing emotions and building user interfaces therefore like brainwave (EEG) signals, mimicry, gestures, eye movement, biometrical signals, speech and written language. Here, some examples out of these are presented to, on the one hand, show the possible benefit of computationally transforming recognized emotions to sensible feedback and, on the other hand, give insight into actual research in that regard. The interested reader may refer to proceedings concerned with affective computing like the ones of *Affective Computing and Intelligent Interaction*.
A very important benefit of computational feedback to given recognized emotions is that individuals can reflect own behaviour if a computer classifies it as not desireable. In that regard, Hoque and Picard [15] conducted experiments with a virtual agent, which recognizes several aspects of talking and behaving during job interviews. The software gives according advices about how to improve verbal and non-verbal expression for these situations. Subjects trained with this systems were rated by uninformed human judges as better performers than other subjects. These experiments clearly show how much persons can benefit from computational feedback to the way we express ourselves. Negative ways of expression like unfriendly mimicry, nervous talking or showing a lack of concentration can computationally be mirrored and thus made conscious to us enabling insights about how others perceive us and what we want to do about it.
These ideas correlate with the ones of the present work, whereas here feedback to recognized emotions is given continuously at the very moment they occur. Additionally, emotion shall be returned to the user without any intentions to influence one's behaviour into a desirable direction, but with the mere claim to make it accessible vividly by means of an according augmentation.
Similar was achieved with iFeel_IM! [16], a prototype for generative affective haptics during emotional communication in text-based online chats. Hereby sentiment analysis is utilized to extract emotion from text of a chat partner. This information is

converted to several kinds of haptical stimuli, which are primarily based on temperature and pressure. Localizations of these stimuli enrich the emotional experience, e.g. a butterfly vibrator at the stomach (love) or a Peltier element chilling down the spine (fear). The work is lacking a detailed user study, but is claiming to be well accepted in preliminary experiments by testers intensivying their chatting experience a lot. From this results positive effects on concentration and intellegibility can be assumed. This scenario illustrates the possible usefullness of affect augmenation, that was in the same way motivated introductorily for this work, too.
Last but not least, Stefan Ehrlich (advisor of this work) could find connections between brain state (EEG) and musical parameters [2]. This was used to show that algorithmically synthesized music can serve as a valid emotional stimulus. In an inverse way he also illustrated closed-loop application scenarios, where EEG signals from a brain computer interface (BCI) are mapped to accordingly perceived music. Test subjects were enabled to "control the affective music feedback by actively changing their way of thinking" [2, p. 5]. These findings highlight the contribution of affective computing to the topic of self-awareness and -reflection. A considerably high potential for application can be seen in medical sciences concerned with mental health, especially treatment of empathic related mental disorders as mentioned in the previous chapter.
Concludingly, computational affective feedback is a new distinct scientific field, which has started to arise in the past years. Yet it is not considered extensively by research, which still has to become aware about the application potential. The previous excerpt shows that not all human expression modalities were used so far to augment them with according affective feedback. To do so with speech and music appears to be a novel and promising approach and hopefully sheds some more light on the usefullness of affective augmentations.

Chapter 3

Overall Processing Pipeline

As initially described, the aim of the present work was defined as creating a musical equivalent of emotion in speech on the fly. The whole process includes speech recording and analysis, emotion classification and music creation – see Figure 3.1. All steps are done continuously, since talking is an ongoing and always changing process, and computationally, i.e. by a software running on a computer.

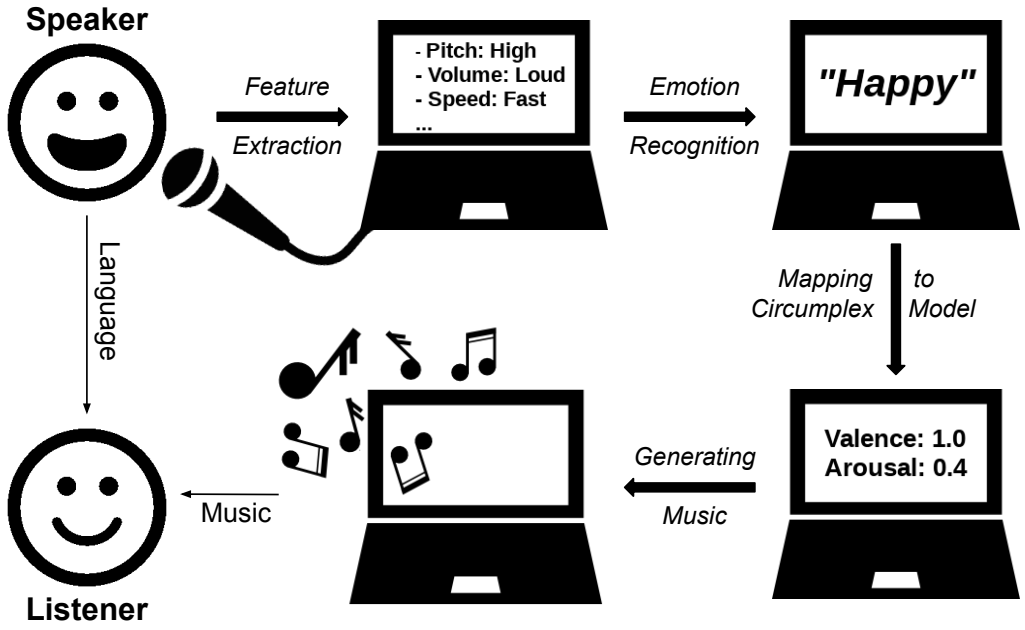

Figure 3.1: Illustration of the processing pipeline from speech over emotion recognition towards music generation.

Continuity in the realm of computation always implies a discrete sampling rate based on which continuously occurring input signals can be recorded, analyzed and processed. In the first step a voice signal from a microfone is analyzed by OpenSmile [17], which returns high dimensional (>1000) numerical vectors (step *Feature*

Extraction). Each dimension thereby represents a prosodic or spectral feature from human speech and the corresponding number the amount of its occurrence.
These vectors serve as basis for a classical classification problem. There are many kinds of categories by which human speech can be classified, for example age, deseases and gender. Here, the interest lies in emotion classification, which is referred to as *Emotion Recognition* in literature – see for example [18]. Thereby an emotion label is assigned to a voice frame of a few seconds. This is technically solved by training a speaker dependent classifier with pre-labeled emotional speech features by means of previously mentioned OpenSmile vectors. As basis for training *Berlin Database of Emotional Speech* [19], in short Berlin-EMO, is used. Lateron, new and unknown speech for the classifier is labeled in accordance to the training data and its probability distributions. For more details about audio feature extraction and emotion classification please refer to the work of Michael Lux.
The emotion labels are transformed to generative music, which is perceived as congruent to the emotion labels. This means the word "anger" must be processed in such a way that computationally generated music is perceived as angry by humans. The same has to hold for all other emotion labels, which are given by the emotion classifier and respectively the spoken training data. As technical solution therefore algorithms exist, which transform two-dimensional numerical values representing emotions from Russel's Circumplex Model to accordingly perceived music. A necessary pre-processing step here is to convert emotion labels to these values called valence and arousal. The concrete implementation of the musical algorithm is described in Chapter 5, whereas the psychological theory of the Circumplex Model and regarding mappings of emotion labels are discussed in Chapter 4.

Chapter 4

Mapping Emotion Labels to Circumplex Coordinates

As denoted beforehand, a speech classifier returns emotion labels representing the emotion which the classifier finds the most likely one at the moment. It was also stated that existing music algorithms require the emotional information as two-dimensional numerical values, namely valence and arousal. The present chapter aims at firstly giving the psychological background knowledge therefore and secondly how emotion labels are mapped to 2D coordinates for the prototype.

4.1 The Circumplex Model of Affect

One way of representing emotions is giving names for these like "anger" or "happyness". Another way of representation is the Circumplex Model of Affect as it was formulated by Russel [20, 21], which stands out due to its significance for the perception of music [22]. As it can be seen on Figure 4.1, it suggests a 2D coordinate system, in which all emotions can be arranged. Their position depends on valence and arousal, whereby valence (x-axis) means if an emotion is experienced as pleasant (positive) or unpleasant (negative). Arousal (y-axis) stands for the degree of stress and movement inherent to the respective feeling.
The intuition behind this way of representation gets clear by a glance on Figure 4.1. Emotions with a high amount of valence are both excitement and contentment, which is why both are situated on the right in the coordinate system. Nonetheless, they differ in their amount of arousal, since excitement implies a higher amount of activity than contentment. Consequently the former has its place at the top right corner whereas the latter is at the bottom right corner. The same with respect to arousal is true for distress and depression, but their amount of valence is strongly negative. That is why these emotions are at the negative left end of the valence spectrum. These four emotions give an imagination of which kind of feelings occur in which quadrant of the coordinate system. On the other hand, arousal, pleasure, sleepiness and misery stand for the positive or negative extremes at each axis.

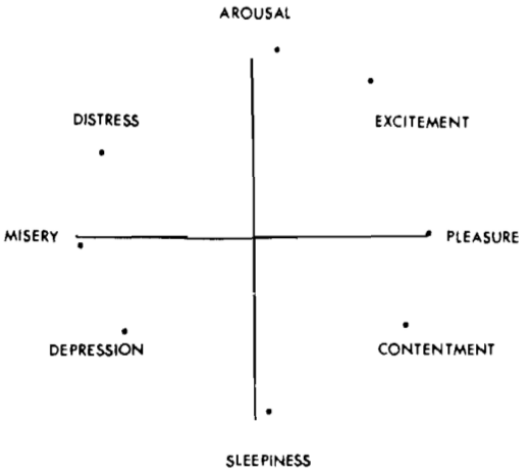

Figure 4.1: Circumplex Model as originally depicted by Russel [20]

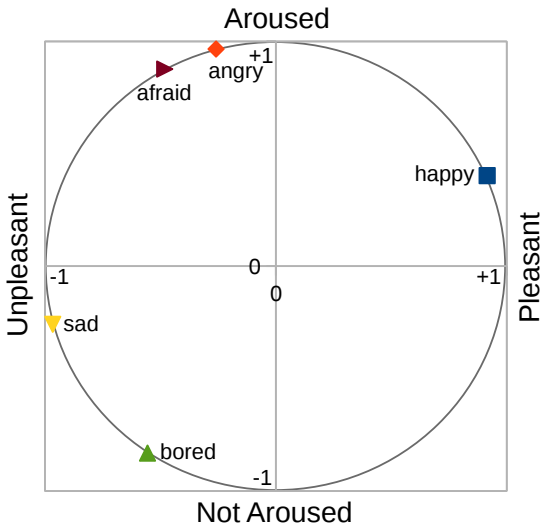

Figure 4.2: Basic emotions from Berlin-EMO [19] mapped to Circumplex Space with circular scaling

These findings are explained in detail by Russel [20]. In his experiments subjects were instructed to place emotion labels in a circle so that "(1) words opposite each other on the circle describe opposite fellings and (2) words closer together on the circle describe feelings that are more similar" [20, p. 4, Circular ordering task]. The overall result of this study is the diagram depicted in Figure 4.2.
Of course the Circumplex Model is not limited to these emotions only. In Figure 4.2 another set of emotions is depicted within Circumplex Space. These are the ones given in Berlin-EMO dataset [19]. In Chapter 4.2 it is explained in detail how exactly this mapping was implemented.

4.2 Mapping Emotion Labels to Circumplex Model

In the previous paragraphs a two-dimensional space was described, in which every emotion has its place according to its respective amount of valence and arousal. Both axes range from -1, i.e. lowest possible amount, to +1, i.e. maximal possible amount. By means of this assumption it can be concluded that emotions are representable by a vector e, for which the following holds:

$$e \in [-1; +1]^2 \tag{4.1}$$

In the following paragraphs a method called *Circular Scaling* is described, by which the exact coordinates in this space for many emotion labels can be calculated. It was firstly described by Russel [20, p. 4 ff.]. He performed a study, in which 36 subjects were given many "stimulus words" describing different emotional kinds of

mental states. The given instruction was to categorize each of these words into one of the eight categories mentioned in the previous paragraphs.
The idea behind this experiment gets clear with a view on Figure 4.1. There the denoted categories are approximately plotted in a circular order with a distance of 45° between each label, whereby the angles for the categories are defined as 0° for pleasure, 45° for contentment, 90° for sleepiness, 135° for depression, 180° for misery, 225° for distress, 270° for arousal and 315° for excitement. The summed up counts of categorizations for each stimulus word can be seen as weights for the angle, which has to be chosen for the respective word. This can be exemplified by the stimulus word "satisfied". Russel therefore stated a categorization count of the following: contentment: 32, pleasure: 3, excitement: 1. Consequently the angle for "satisfied" would be calculated with the formula of the Arithmetic Mean like this:

$$\frac{32 \cdot 45° + 3 \cdot 0° - 1 \cdot 45°}{36} \approx 39° \tag{4.2}$$

For excitement seemingly another degree (-45°) was chosen than the one stated earlier (315°). Actually -45° and 315° is the same angle due to modulo 360° invariance. It is necessary for a correct mean value to measure the degree starting from the category with the maximum categorization count, which in this case is contentment with a count of 32 and an angle of 45°. If 315° would be taken for excitement the arithmetic mean would be distorted heavily into the exact opposite direction. In this case this would be negative valence and high arousal, obviously not suitable for the word "satisfied".
In this work the degrees are calculated for stimulus words taken from Berlin-EMO dataset, which are anger, boredom, anxiety/fear, happiness and sadness. A two-dimensional coordinate for each degree was calculated by taking the intercept point between the unit circle and a line from the center in the direction of the respective angle – this is nothing other than the cosine (x) and sine (y) of the angle. The resulting coordinates for the basis emotions from Berlin-EMO dataset are shown in Table 4.1.

Emotion	happy	angry	sad	bored	afraid
Coordinate	(0.92,0.40)	(-0.26,0.97)	(-0.97,-0.26)	(-0.56,-0.83)	(-0.48,0.88)

Table 4.1: The resulting coordinates within Circumplex Model for the emotions from Berlin-EMO dataset [19] as they were used in this work

The performed calculation does not scale down the distance from the origin, as the original procedure from Russel does. This decision was made to have more extreme valence and arousal values, which leads the music generation to be more emotional and thus more supporting and distinguishable. Since the chosen emotion labels from Berlin-EMO are also extreme emotions within the Circumplex Model, this can be seen an legetimate. Nonetheless, this behaviour can be questioned if more neutral emotions are of interest, what does not apply here – the stimulus word "disgust"

from Berlin-EMO was omitted out of the mentioned reasons. It is advised to refer to Russel's section about P values if this behaviour is not wanted [20, p. 6].

Chapter 5

An Object-Oriented Generative Music Algorithm

So far a processing pipeline was discussed, that recognizes emotion from speech and transforms emotion labels to a two-dimensional numerical representation of the Circumplex Model. This chapter shows how this information is computationally transformed to generative music in such a way, that listeners assess the music equally to the intended input emotion. It shall be emphasized here that the produced music is in no way based upon recorded audio signals. Instead it is generated based on dynamic and probabilistic principles making it unique at every moment in time when it is generated.

5.1 Related Work

The Circumplex Model was introduced into musicology by Gabrielsson & Lindström [23]. They 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 [24, 25]. Similar results could be achieved by manipulating score and performance features of existing music pieces so, that the expressed emotion changed in an intended way [26]. In view of that, existing software from research can generate music with previously determined affect.
Against this background, the advisor of this work Stefan Ehrlich [2] implemented an own procedural music algorithm in Matlab called *Composer*. Its principles are explained in Section 5.2. Based on that, a new reimplementation was done in Python called *PyComposer* – see Section 5.3. The aim was to show a reasonable object-oriented approach for these kind of musical algorithms.

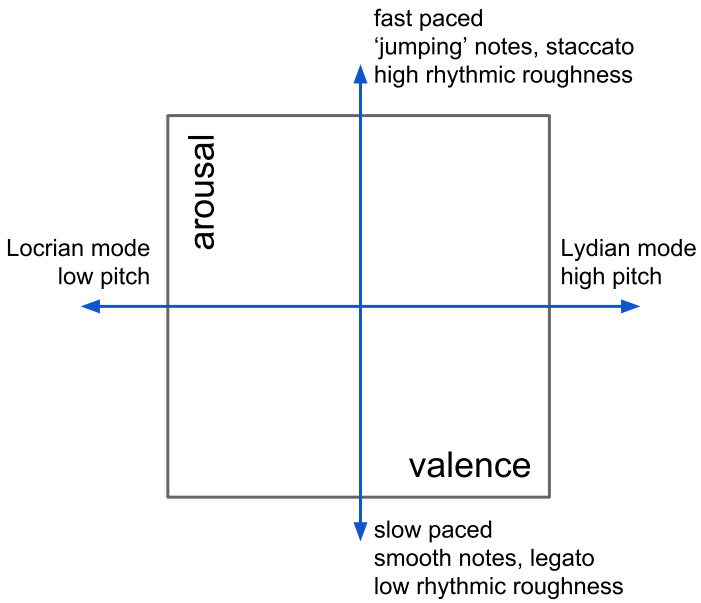

Figure 5.1: Relations between valence & arousal and musical rules: While valence is correlated especially with the mood of modes, arousal is connected with aspects of speed and stress. Taken from [2].

5.2 Circumplex Model & Transformational Music Rules

The following lines will lay out how valence and arousal are correlated with musical rules. Since this knowledge is extensively explained in free accessible referred research, which was not the primary concern of this work, several details about it are left out to give merely an overview of the used techniques.
In simplified terms, valence corresponds to the mood of a scale. For example minor tonalities are considered to sound sad, which implies correlation to lower valence values. The opposite holds for major tonalities. In this context Gomez [22] found an ordering of modes according to the amount of valence they evoke in listeners sensation. Technically the ordering is a circle of fourths starting at the seventh of an Ionian scale. In C Major this would correspond to B and the mode would be Locrian. The next would be E Phrygian and so forth. In this way minor, sad and thus low valence kind of scales are at the beginning of the ordering, while major, bright and high valence kind of scales are at the end of the spectrum. In dependence of the actual valence value a fitting mode can be chosen. Based on this tonality selection a repeating chord progression is arranged over the degrees 1–4–5–1. The sensations of the intended moods can additionally be supported by choosing tones with a lower pitch for low valence modes and higher tones for high valence modes.

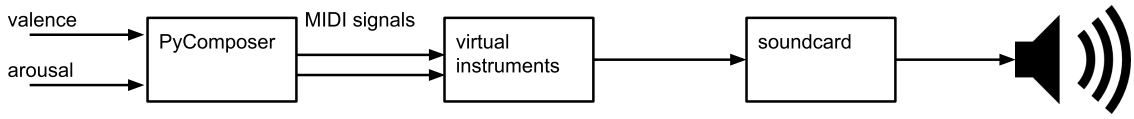

Figure 5.2: Processing of valence & arousal to music: Implementation is primarily concerned with PyComposer, which transforms the value pairs to MIDI signals. These are processed to audio signals by virtual MIDI instruments and the sound card. Taken from [2].

Doing this is possible via "voicings, which define the pitch register where the notes are going to be played" [2], see implementation details lateron.
Arousal is connected to musical features related *to the way* the tones are played. This includes among others tempo (slow/fast), pitch range (narrow/wide), articulation (legato/staccato) and rhythmic roughness (low/high) – see Figure 5.1.

5.3 PyComposer – An Object-Oriented Music Generator

When emotions from speech are represented by 2D positions in the Circumplex Model, these values are passed to the incorporated music algorithm conveying them to accordingly perceived music. In the first step a MIDI pattern is generated by a state machine called *PyComposer*, which is the central part of the music related implementation of this work. The MIDI signals are then translated into sound by virtual instruments from external software, for example ProTools or Qsynth. The explained dataflow is depictured in Figure 5.2.
In the following the class structure, algorithmic details and the operation mode of PyComposer are described with particular focus on object orientation, which was added in this project.
An overview about the class structure is given in Figure 5.3. Beginning from valence and arousal, these values are saved and hold by the State object enabling thread-safe read and write access to these values. This is a critical part, since these values are changed and read by two different threads: recognizing emotion (write) on the one hand and the music generation (read) on the other. While the former process is not shown in Figure 5.3, the latter thread is defined by the StatePlayer class. By starting its play() method it reads the valence and arousal values and creates music for one bar. This procedure is repeated until stop() is called – see Listing 5.1 as code example.

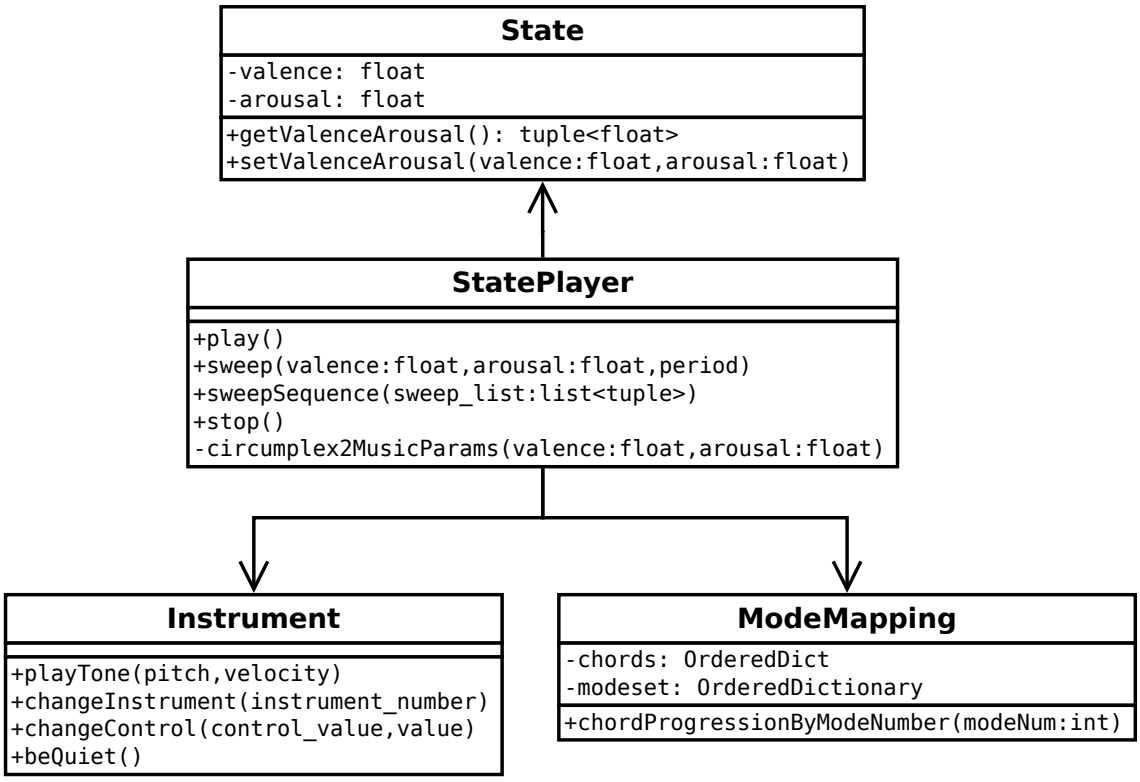

Figure 5.3: UML class diagram of PyComposer, an object-oriented state machine for generating music based on valence and arousal.

Listing 5.1: Example of how to start generating music with PyComposer: After initialization (line 1–2) a valence arousal value pair is set (line 3). According music starts playing (line 4). Then a smooth transition (sweep) to another musical mood takes place within 30 seconds (line 5). Sweeps can be stacked and executed one after another (line 6).

```
1 from PyComposer.StatePlayer import StatePlayer
2 with StatePlayer() as player:
3     player.setValenceArousal(1,1)
4     player.play()
5     player.sweep(-0.1, -0.8, 30)
6     player.sweepSequence([(-1,-1,20),(-1,1,30),(1,1,40)])
7     player.stop()
```

A key element thereby is the private static circumplex2MusicParams() method called from within play(). It transforms valence and arousal to several musical structural parameters like the mode number, tempo, stroke velocities and many more. The mode number is passed to ModeMapping.chordProgressionByModeNumber(). Its result is a sorted dictionary of the chord sequence. The format used is exemplified in Listing 5.2.

Listing 5.2: Example of how chords and modes are defined in source code: 4-note chords are saved based on their root within the tonality – pitches are given in integers according to MIDI standard. Modes are stored based on their names and include chords of the progression 1-4-5-1.

```
1  chords, modes = OrderedDict(), OrderedDict()
2
3  chords['tonic'] = [60, 64, 55, 59]  # C, E, G, H -> C Maj7
4  # do the same for supertonic, mediant & all other degrees
5
6  modes['ionian'] = [
7      chords['tonic'],
8      chords['subdominant'],
9      chords['dominant'],
10     chords['tonic'] ]
11 # do the same for phrygian, aeolian & all other modes
```

From the chord sequence the actual chord is picked by StatePlayer and its tones given as MIDI pitches are passed to Instrument.playTone(). Additionally velocity is passed including the information how strong the tone has to be played. That is calculated beforehand by circumplex2MusicParams(). Which MIDI instrument is used can be controlled by changeInstrument(). There the instrument number according to MIDI standard is passed. From these explanations and Listing 5.3 it

gets clear, that the Instrument class is not more than a programmer friendly MIDI instrument interface.

Listing 5.3: Example of how to use a MIDI instrument: After initialization (line 1–2) a piano is chosen as MIDI instrument (line 3). Then a gentle C-tone from the piano occurs (line 4). In addition a contrabass is added on channel 2 (line 5). This is heard as an loud E (line 6).

```
1 from PyComposer.Instrument import Instrument
2 with Instrument() as i:
3     i.changeInstrument(1)
4     i.playTone(60,50)
5     i.changeInstrument(44,[2])
6     i.playTone(64,70,[2])
```

5.4 Remarks on MIDI Instrumentation

In this work besides object orientation another remarkable upgrade has been made to affective music generation, that has already been proven to work in [27]. The algorithm chooses MIDI instruments dependent on valence and arousal. For example the bass voice is distorted when valence is low and arousal is high, creating a more gloomy atmosphere. For low arousal values instead of normal bass tones a contrabass instrumentation is chosen making the accompaniment more soft and relaxing. To the author's best knowledge this was the first successfull attempt of utilizing MIDI instrumentation for affective music algorithms, especially in view of the positive results of Chapter 6.

Chapter 6

Evaluation of the Prototype

As denoted, the aim of the prototype is to support expressivity of speech by generating according music on the fly. Success in this context is when affect from speech and generated music being perceived as equal or similar. If that condition is met, emotional information from voice is augmented musically by the proposed approach. To show this is possible with the prototype, an evaluation in terms of a user study is mandatory. If subjects assess the emotion from music equally or similarly as from voice, the mentioned requirement can be argued to be met. This chapter exposes such an evaluation as it was conducted for this work.

6.1 Preparation

Emotional speech recordings are conducted for each emotion from the Berlin-EMO dataset seperately. This means one sound file is created containing only angry voice samples, the next one only happy and so forth. Afterwards, for each of these voice files music is generated by the prototype and recorded, too. As a result, there exists one music and one speech recording for each emotion. From this, two classes of sound files are created: congruent and non-congruent. For the former, speech is overdubbed with fitting music[1]. For the latter speech is overdubbed with music randomly chosen out of the not-fitting pieces. For each class, i.e. congruent and non-congruent, seven overdubbed sound files are created, which makes 14 in total.

6.2 Survey

These audio files were played to subjects in a random order. As subjects 27 people took part – additional information like gender, age and further background was not acquired. After each file they had to answer the question if the emotion from speech was similar to the emotion in the background music. Discrete values ranging

[1] In this context *fitting music* denotes music, that was generated to the respective emotional speech by the prototype.

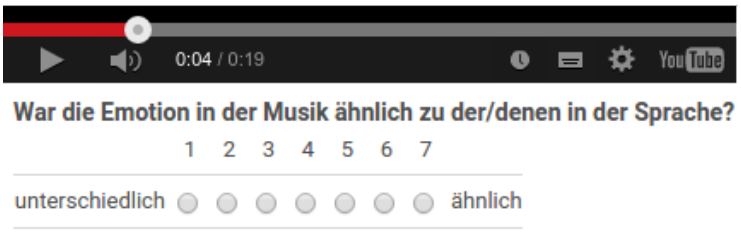

Figure 6.1: Example question (screenshot) as it appeared in the actual questionnaire

from 1 to 7 with step size 1 could be given as answer according to the style of Semantic Differential. Only one value could be stated per question. In Figure 6.1 the appearance of a question with corresponding sound playback is shown.

6.3 Results

To statistically evaluate the questionnaire data, the following computation steps were performed. The answers of each individual were averaged by class, i.e. the ratings of all congruent sound files were averaged for each subject and the same for all non-congruent. This is done, because the distribution of the average rating per class for one person is of interest. The results were two values per individual: congruent average and non-congruent average.
The results are shown in Figure 6.2. For this illustration all averages were rounded to 0.5 precision and plotted in the style of a histogram. Clear differences between both classes are visible: The congruent sound files (blue bars) were rated as more congruent, which makes them appear more at the right of the diagram compared to the non-congruent sound files (orange bars), which are positioned rather on the left and are thus more rated as non-congruent.
Even though this can be stated as positive with respect to the initially mentioned requirements, it is not enough in terms of statistical significance. Therefore further statistical processing is necessary. The relevant question here is, if the two distributions of the average rating per class, i.e. congruent and non-congruent, are significantly different from each other. If this is the case, then it is very likely that the distributions are in fact different. This in turn would lead to the conclusion that, due to the obviously different mean values, the music of the prototype is generally perceived as emotionally more similar to speech as completely random generated music.
To statistically compare the two distributions, it is at first necessary to inspect if both are normal distributed. Therefore we choose Shapiro-Wilk and Anderson-Darling tests, since both are proven to yield the best results for the stated problem [28], especially compared to Kolmogorov-Smirnov and Lilliefors. Both tests are valid with respect to the sample size, which has to be at least 8 for Anderson–

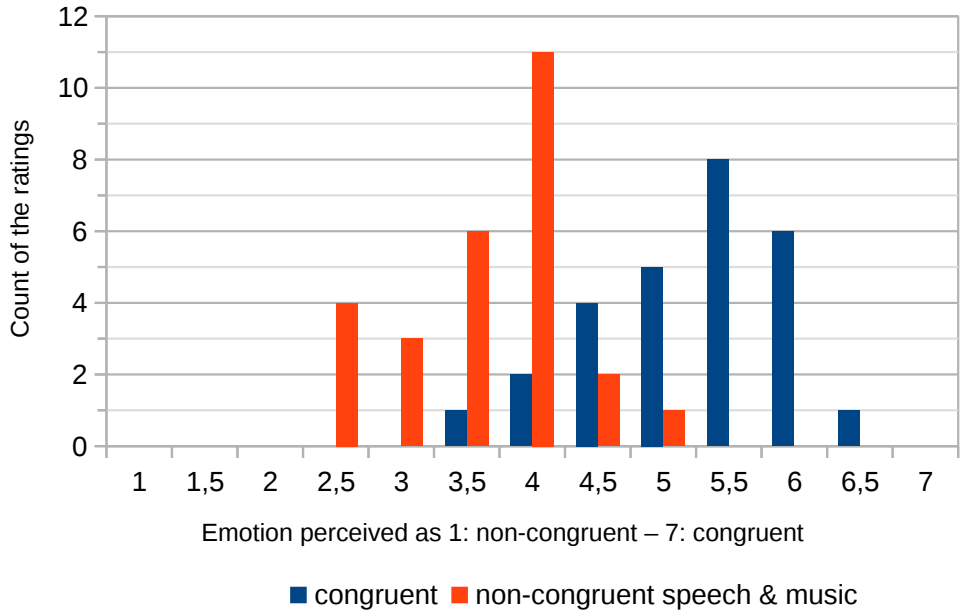

Figure 6.2: Rating histogram of congruent speech + music and non-congruent: The former (dark blue, mean 5.3) was perceived as more congruent than the latter (orange, mean 3.6).

	Congruent		Non–Congruent	
	h	p	h	p
Anderson-Darling	0	0.09	0	0.11
Shapiro-Wilk	0	0.10	0	0.21

Table 6.1: P-Values and regarding test decisions (h=0 means no rejection based on 5% significance level) of the ratings of congruent and non-congruent speech & music

Darling and 3 for Shapiro–Wilk. Their results are shown in Table 6.1. From there it can be seen, that for the ratings of both congruent and non-congruent sound files the null hypothesis cannot be rejected, that they follow a normal distribution (significance level: 5%). Even though this is true for both applied tests, it does not mean, that normal distributions lie at hand. Nonetheless, this can be taken as a reasonable argument to proceed as if this would be the case. Consequently, two-sample T-Test is used to see if both distributions – assuming they are normal – are equally distributed. T-Tests clearly rejects the null hypothesis that both distributions have equal means by a p-value of $3 \cdot 10^{-11}$, whereby nothing is assumed about the variances. So, under the assumption of normal distributions, it can be said both distributions are significantly different. For the sceptic reader it shall be noted, that even under the assumption of arbitrary, but identical distributions, Wilcoxon signed-rank test rejects the hypothesis of same distributions clearly by a p-value of $8 \cdot 10^{-06}$. Concludingly, distributions for the ratings of congruent and non-congruent sound files are different with a very high significance. In view of Figure 6.2 and different mean values, 5.3 for congruent and 3.6 for non-congruent sound files, speech with congruently overdubbed music is rather perceived as such as arbitrarily overdubbed speech.

Chapter 7

Conclusion

In this work a software prototype was presented, which recognizes emotion from prosody in speech and transforms this information into generative music. In view of the evaluation results, it can be stated that this approach is reasonable, since the emotion from generated music is perceived as similar to the emotion from speech. These findings are remarkable, because they are unique so far to the best knowledge of the authors. Not only the invention itself to augment emotion from speech with computer music can be considered as highly innovative input for research and and multimedia industry as already mentioned in Chapter 1. Also the concept of how this was realized, i.e. the emotion processing pipeline as seen on Figure 3.1, has the potential to serve as generic pattern to augment music wherever emotions from individuals are computationally recognized, .

With respect to concrete application of the explained prototype for further research or industry products, the necessary steps therefore appear manageable. Realtime music production to speech is already possible and has been exemplified for a few demonstration sound files [29]. Several drawbacks still exist in that regard. For example the utilized emotion recognition from speech is neither speaker nor channel invariant – see the work of coauthor Michael Lux. Nonetheless, for this purpose professional commercial industry solutions already exist, for example from audEERING UG [30]. By taking advantage of this, the emotion recognition module can simply be exchanged by a professional solution enabling use in everyday conditions with headphones, to musically augment emotions from people talking to oneself for example. For use with speakers, e.g. to accompany artistical verbal expression, the problem needs to be faced that the generated music will in turn influence emotion recognition, since it affects the incoming audio signal. Techniques like background noise reduction or beam forming may solve this problem. Furthermore machine learning based speech enhancement specifically designed for musical accompaniment can be considered as most promising and state of the art. One may refer to a commercial professional solution from audEERING here, too, or regarding literature like [31].

List of Figures

Bibliography

[1] J. K. Baghel. Audio-based characterization of conversations. Master's thesis, Technische Universität München, Institut für Informatik, 2014.

[2] S. Ehrlich. Eeg emotion recognition and affective music bci. Master's thesis, Technische Universität München, 2013.

[3] Gerhard Johann Hagerer, Michael Lux, Stefan Ehrlich, and Gordon Cheng. Augmenting affect from speech with generative music. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 977–982. ACM, 2015.

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

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

[6] E. D. Ross. Affective prosody and the aprosodias. 2000.

[7] Marjolijn H., Rene S., Pijnenborg, K. & M., Aleman, A. Impaired recognition and expression of emotional prosody in schizophrenia: Review and meta-analysis. *Schizophrenia Research*, 96(1 – 3):135 – 145, 2007.

[8] M. Ghaziuddin. Defining the behavioral phenotype of asperger syndrome. *Journal of Autism and Developmental Disorders*, 38(1):138–142, 2008.

[9] 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):788–793, 2007.

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

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

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

[13] Galizio, M., Hendrick, C. Effect of musical accompaniment on attitude: The guitar as a prop for persuasion. *Journal of Applied Social Psychology*, 2(4):350–359, 1972.

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

[15] M. Hoque and R.W. Picard. Automated coach to practice conversations. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 703–704, Sept 2013.

[16] D. Tsetserukou, A. Neviarouskaya, H. Prendinger, N. Kawakami, and S. Tachi. Affective haptics in emotional communication. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pages 1–6, Sept 2009.

[17] Eyben, F., Weninger, F., Groš, F., Schuller, B. Recent developments in opensmile, the munich open-source multimedia feature extractor. pages 835–838, October 2013.

[18] Theodoros Kostoulas, Todor Ganchev, Alexandros Lazaridis, and Nikos Fakotakis. Enhancing emotion recognition from speech through feature selection. In *Text, speech and dialogue*, pages 338–344. Springer, 2010.

[19] Bartels, A., Rolfes, M., Burkhardt, F., Technical University Berlin. Berlin database of emotional speech, requested on 20/11/2014 11:30am.

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

[21] Russell, J. A., Weiss, A., Mendelsohn, G. A. Affect grid: A single-item scale of pleasure and arousal. 1989.

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

[23] Gabrielsson, A., Lindström, E. The influence of musical structure on emotional expression. 2001.

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

[25] 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.

[26] 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*, 34:41, 2010.

[27] Georg Groh Klügel Niklas, Gerhard Hagerer. Fuguegenerator - collaborative melody composition based on a generative approach for conveying emotion in music. In *Joint International Conference Music Computing (ICMC) and Sound and Music Computing (SMC)*, pages 286–293. National and Kapodistrian University of Athens, 2014.

[28] Nornadiah Mohd Razali and Yap Bee Wah. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33, 2011.

[29] Hagerer, Gerhard Johann. Papers and thesises of gerhard hagerer, requested on 08/09/2015 10:49am.

[30] audEERING UG. audEERING – intelligent Audio Engineering, Website, requested on 08/09/2015 11:29am.

[31] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller. Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr. In *Latent Variable Analysis and Signal Separation*, pages 91–99. Springer, 2015.