ZUR WAHRNEHMUNG VON MANIPULIERTEM WEINEN ALS LACHEN

Jürgen Trouvain

Phonetik, Universität des Saarlandes
trouvain@coli.uni-saarland.de

Abstract: Prototypische von Schauspielern produzierte non-verbale Vokalisierungen zum Ausdruck von "glücklich" und "traurig" ähneln sich stark in ihrer akustischen Ausprägungen, weswegen sie zuweilen von Hörern verwechselt werden. Im vorliegenden Aufsatz wird von Schauspielern produziertes Lachen und Weinen phonetisch verglichen und ein Pilottest vorgestellt, bei dem Manipulationen von Weinen zur Wahrnehmung als Lachen führt.

1 Einführung

Affektivität im Sprachsignal spielt eine stetig zunehmende Rolle sowohl in der automatischen Spracherkennung (vgl. [8]) als auch in der Sprachsynthese (vgl. [4]). Zu den mitunter als stark empfundenen affektiven Zuständen, die durch Vokalisierungen geäußert werden, gehören Lachen und Weinen, wobei Lachen meist (aber nicht immer) mit einer positiven Bewertung einhergeht und Weinen meist (ebenfalls nicht immer) mit einer negativen. Von einer akustischen Perspektive aus betrachtet ist es interessant, dass stark ausgeprägte Vokalisierungen beider Affektkategorien sich sehr ähnlich sind. Daher ist es nicht verwunderlich, dass bestimmte Formen von Lachen mit Weinen von Hörern verwechselt werden und umgekehrt bestimmte Formen von Weinen fälschlicherweise als Lachen interpretiert werden.

Für die vorliegende Pilotstudie wurden von Schauspielern produzierte und als prototypisch eingeschätzten Affektäußerungen von "traurig" und "glücklich" phonetisch analysiert. Anschließend wurden auf diesen Äußerungen aufbauend Stimuli für einen Wahrnehmungstest erzeugt.

2 Phonetische Analyse prototypischer Vokalisierungen von Lachen und Weinen

2.1 Datenbank

Grundlage für die vorliegende phonetische Analyse ist die Montreal Affective Voice Database [3]. Für die Affektäußerungen haben 10 Schauspieler (5 männlich, 5 weiblich) Emotionen der Kategorien Wut, Ekel, Angst, Schmerz, Trauer, Freude, Überraschung, Vergnügen sowie "neutral" produziert. Von allen Äußerungsversuchen eines jeden Schauspielers wurden von Hörern für jede Emotionskategorie als prototypisch eingeschätzten Version ausgewählt. Aus diesen besten Exemplaren werden die Äußerungen von "sadness" und "happiness" hier genauer untersucht (insgesamt 20 Signale).

2.2 Akustische Ähnlichkeiten

2.2.1 Rhythmische Ähnlichkeit

19 der 20 Äußerungen enthalten eine staccatoartige rhythmische Struktur, die Wiederholungen einer Silbe ähneln (für "prototypisches Lachen" vgl. [10, 9]). Neun der zehn Schauspieler zeigen daher vergleichbare rhythmische Strukturen für Lachen und Weinen (vgl. Abbildung 1). Eine Schauspielerin ("53" in der Datenbank, identisch mit Sprecher 5 in den

Abbildungen) zeigt für ihre Darstellung des Weinens keine "Silbenstruktur", sondern eine unregelmäßige Folge heterogener schluchzartiger Laute. Diese Art von Vokalisierung kommt partiell auch bei den anderen Schauspielern als Onset von Lachen und Weinen vor (vgl. S2 und S10 in Abb. 1).

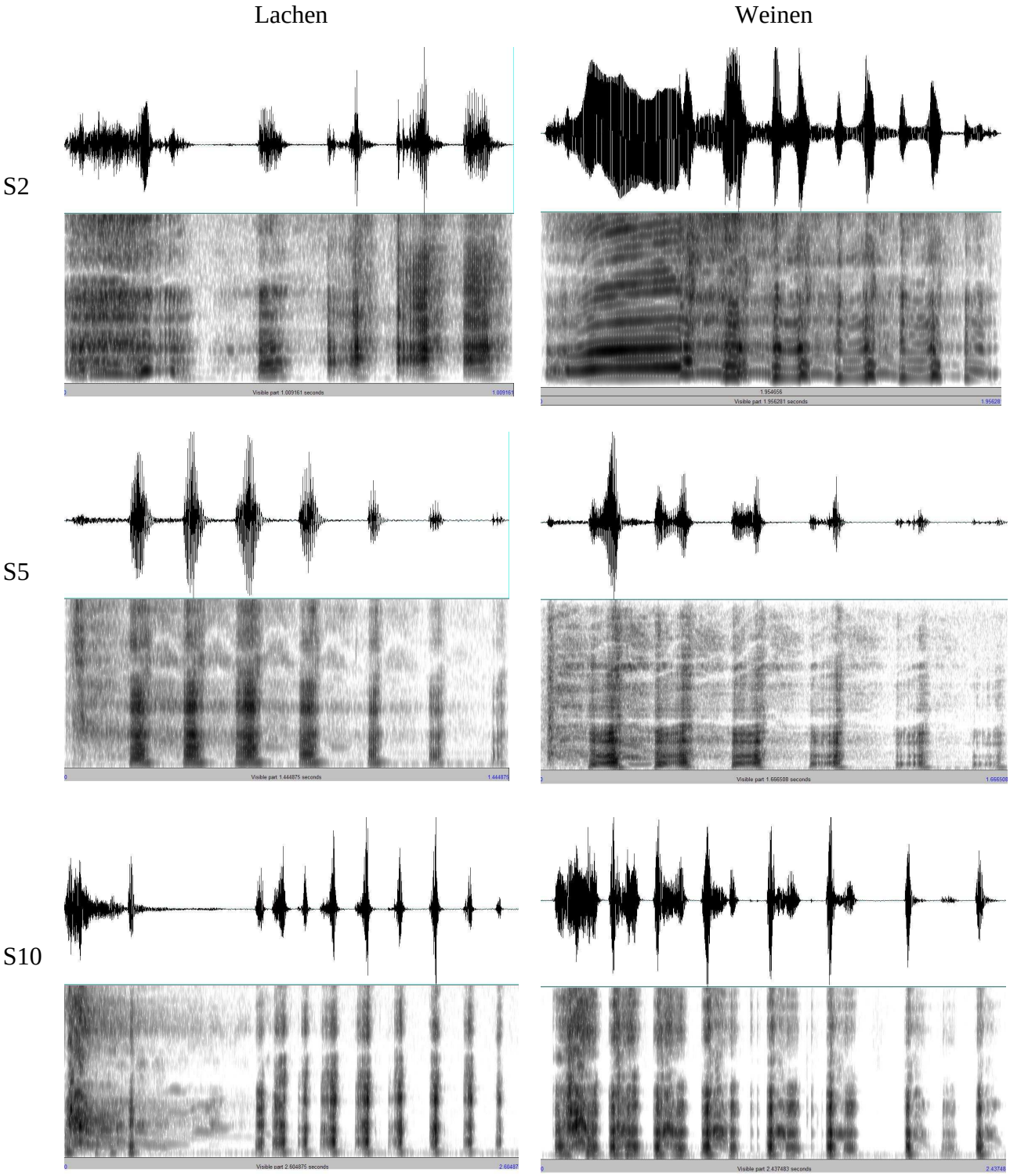


Abbildung 1 – Zeitsignale und Spektrogramme (0-8 kHz) von drei Schauspielern (oben: S2 ("46" in [3], Mitte: S5 ("42" in [3]), unten: S10 ("61" in [3])) aus der Montreal Affective Voice Database für "happiness" (jeweils links) und für "sadness" (jeweils rechts).

2.2.2 Tempo und Dauer

Tempo im Sinne von Silben pro Sekunde artikulierter Sprache kann bei non-verbalen Äußerungen natürlich nicht gelten. Was beim Lachen dem Korrelat einer Silbe entspräche wird oft als "Call" bezeichnet [2, 9]. Daher kann das Tempo als "Call"-Rate wiedergegeben werden. Im Vergleich zeigen alle Schauspieler für das Lachen ein höheres Tempo als für Weinen (vgl. Abb. 2). Der Durchschnittswert für Lachen beträgt 5,3 "Calls" pro Sekunde, derjenige für Weinen 3,4.

Eine Betrachtung der jeweiligen Gesamtdauer aller 20 Äußerungen zeigt, dass die Lach-Äußerungen in 8 von 10 Fällen kürzer sind als diejenigen des Weinens: im Schnitt 1,446 sec zu 2,229 sec des Weinens. In den zwei Ausnahmefällen liegen die Dauern beider Affekt-signale nahe beieinander.

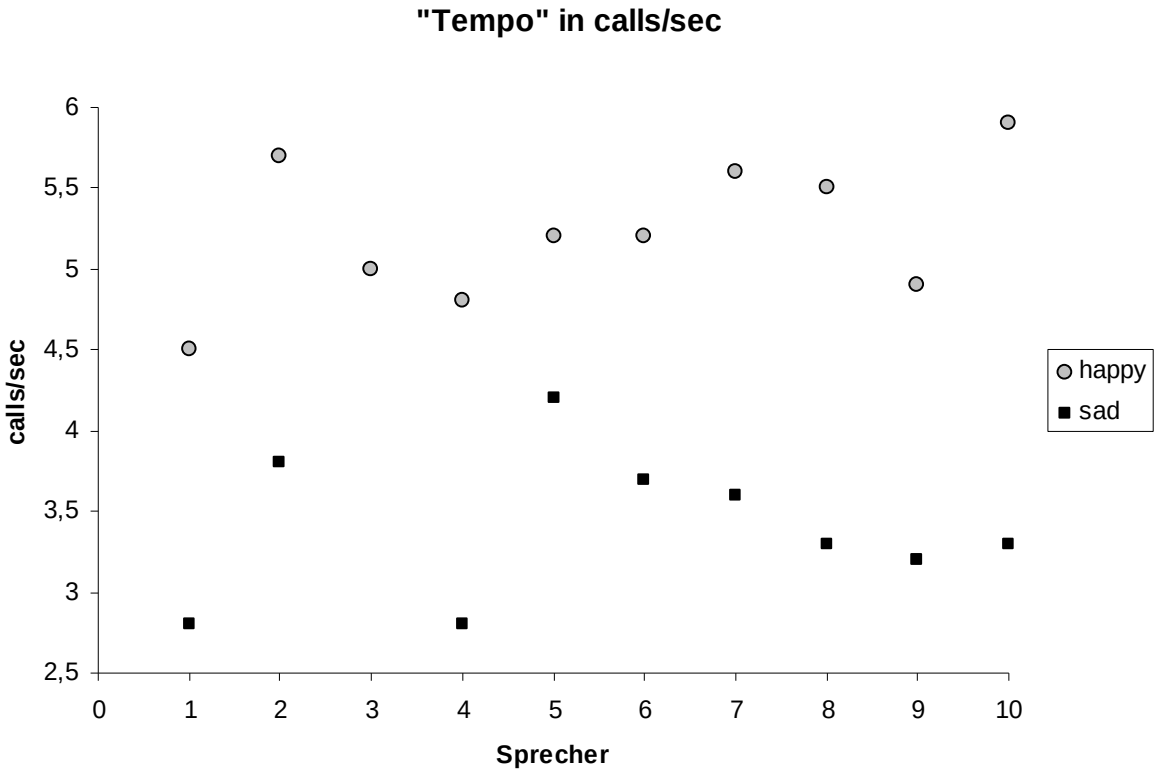


Abbildung 2 – Tempo in "Calls" pro Sekunde der 19 vergleichbaren Vokalisierungen für "happiness" und für "sadness" aus der Montreal Affective Voice Database.

2.2.3 Grundfrequenz

Für beide Affekt-Vokalisierungen lassen sich stark erhöhte Grundfrequenzwerte im Vergleich zu "neutral" feststellen. Für die untersuchten Männerstimmen liegen die F0-Werte teilweise über 500 Hz und für die Frauenstimmen teilweise über 700 Hz. Eine strikte Trennung von Lachen und Weinen wie beim Tempo gibt es bei der Grundfrequenz nicht. Es gibt eine Tendenz, dass Weinen höhere F0-Werte aufzeigt (vgl. Abb. 3). Nur jeweils ein Mann und eine Frau produzieren ihr Lachen mit einer höheren F0 als ihre Äußerung für "traurig".

Bezüglich der Grundfrequenz gibt es große inter-individuelle Unterschiede, was man in Abbildung 3 gut erkennen kann, wenn man die beiden Sprecherinnen 1 und 3 sowie die Sprecher 7 und 9 (beide männlich) vergleicht.

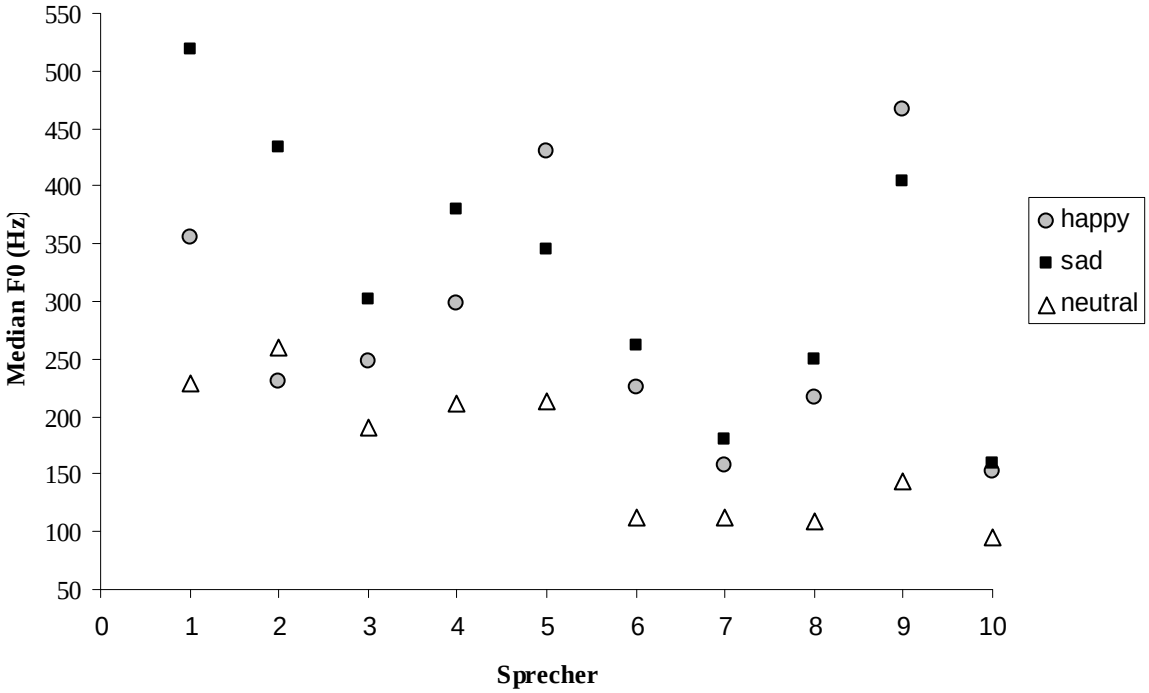


Abbildung 3 – Grundfrequenz als Medianwerte (in Hz) der Vokalisierungen für "happiness", "sadness" und "neutral" aus der Montreal Affective Voice Database. Sprecher 1-5 sind weiblich, Sprecher 6-10 männlich.

2.2.4 Intensität

Betrachtet man die Medianwerte der Intensität in [3], so kann man keine eindeutige Unterscheidung zwischen Weinen und Lachen festhalten. Einige Schauspieler nutzen Intensität in sehr starker Weise zur Differenzierung von Lachen und Weinen, andere kaum; einige zeigen höhere dB-Werte für Lachen, andere für das Weinen.

Bezüglich der Intensität (oft auch bezüglich der Grundfrequenz) lässt sich aber für die meisten "Sprecher" einen Deklinationseffekt beobachten, der sich in einem Decrescendo äußert: eine Reduktion der Intensität beim Weinen über die gesamte Äußerung hinweg und beim Lachen zumeist nur über die letzten beiden "Silben".

2.2.5 "Vokale"

Beim Lachen und beim Weinen kann es zu vokal-artigen Lauten kommen. Eine Betrachtung der "Vokale" zeigt ein weiteres wichtiges Unterscheidungsmerkmal: beim Lachen gibt es häufig ein einfaches stimmhaftes vokalisches Signal mit *einem* Gipfel, beim Weinen hingegen sind sehr häufig *zwei* Intensitätsgipfel am Anfang und am Ende des "Vokals" erkennbar (vgl. Abb. 4 rechts). Auch wenn diese Art von Binnenstruktur vereinzelt auch beim Lachen vorkommt (wie in Abb. 4 links), so scheint sie eher typisch für das Weinen, wobei es hier eine größere inter-individuelle Variation gibt. Zudem beginnen die Weinäußerungen häufig mit einem sehr gedehnten Vokal.

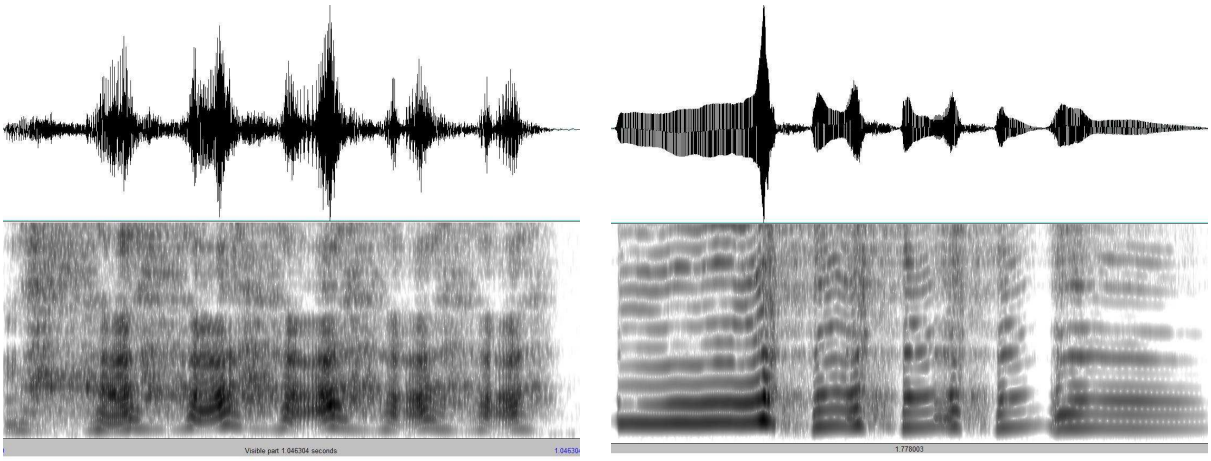


Abbildung 4 – Zeitsignal und Spektrogramm eines typischen Beispiels für doppelte Intensitätsgipfel der "Vokale", sowohl bei "happy" (Beispiel links, S4 ("58" in [3]) als auch bei "sad" (Beispiel rechts, S1 ("45" in [3])).

3 Signal-Manipulationen

Die Autoren der Montrealer Datenbank sind sich bei der Diskussion über die Nutzung der Affektäußerungen als Stimuli in Wahrnehmungstests der Verwechslungsanfälligkeit von Lachen und Weinen bewusst: "Editing the cries […] in an attempt to make them shorter and more compatible with the duration of other sounds could potentially affect their naturalness and would probably lead to important confusions with the laughs." [3: 537]

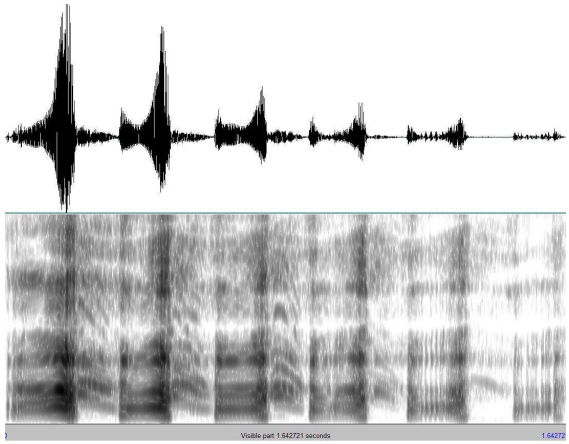
Eine andere Art von Manipulation der temporalen Struktur führt zu ähnlichen Gedanken. Schließlich drängt sich bei der vergleichenden Betrachtung von Lachen und Weinen die folgende Frage auf: Reicht es aus die rhythmische Struktur des Weinens zu ändern, um ein Lachen beim Hörer hervorzurufen? Dazu wurden Manipulationen an den "traurigen" Audiosignalen für einen informellen Pilottest vorgenommen.

Die "traurigen" Äußerungen werden auf zwei Arten bearbeitet, so dass die neuen Signale die identische Dauer zeigen, gleich nach welcher Methode manipuliert wird (vgl. Abb. 5):

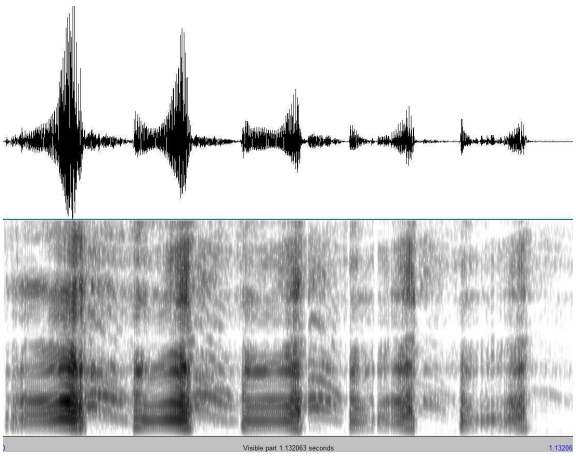
1. Eine globale Tempo-Anpassung: Es wird eine lineare Beschleunigung vorgenommen, so dass das "Tempo" des "traurigen" Signals dem Tempo des "glücklichen" Signals des jeweiligen Sprechers entspricht.

2. Lokale Tempo-Anpassungen: die Dauer der "Konsonanten" und "Vokale" sollen durch Entfernen von Signalanteilen angepasst werden. Der intensivere der beiden Gipfel des "Vokals" wird für jede "Silbe" behalten, entfernt werden die Signalanteile um diesen vokalischen Gipfel, vor allem diejenigen mit wenig akustischer Information. Es wird solange entfernt bis die Dauer des "Konsonanten" im manipulierten Signal die Dauer des "Konsonanten" im ursprünglichen Lach-Signal des jeweiligen Sprechers erreicht wird. Dies führt letztendlich auch zu einer Tempoanpassung des gesamten Signals.

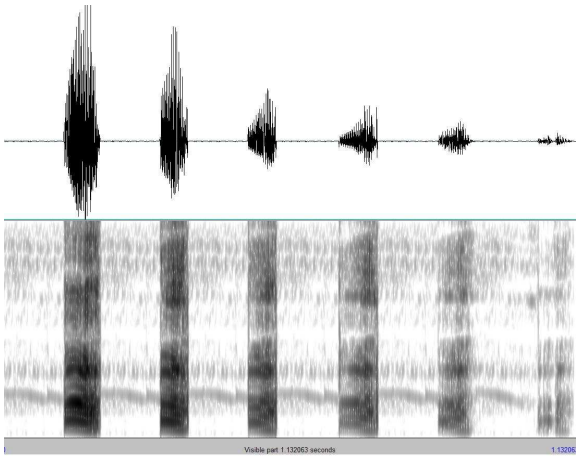
original "sad"



manipuliert nach Methode 1



manipuliert nach Methode 2



original "happy"

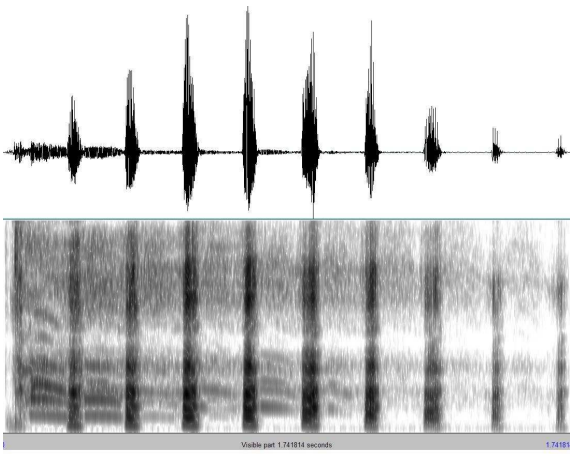


Abbildung 5 – Wellenformen und Spektrogramme der Affektäußerungen von Sprecher 6 ("6" in [3]): oben: Original-Version "sad", unten: Original-Version "happy", Mitte links: nach Methode 1 manipulierte Version von "traurig" mit Ziel "happy", Mitte rechts: nach Methode 1 manipulierte Version von "traurig" mit Ziel "happy".

Als Vorstufe für noch durchzuführende Perzeptionsexperimente wurden in einem informellen Pilottest vier Versuchspersonen in randomisierter Reihenfolge die manipulierten Signale vorgespielt, die auf 9 der zu verwertenden "traurigen" Äußerungen basieren. Nach jedem Stimulus sollten die Versuchspersonen frei antworten, was sie hörten.

Wie zu erwarten führt simples Beschleunigen der Weinen-Signale (Methode 1) nicht zur erhofften Wirkung, dass die Zuhörer ein Lachen wahrnehmen. Eine Reduktion auf einen Gipfel des "Vokals" mit enstprechender rhythmischer Anpassung (Methode 2) hingegen lässt bei manchen Stimuli die Zuhörer tatsächlich glauben, dass es sich hierbei um ein Lachen handelt (z.B. bei dem Signal in Abb 5 Mitte links). Allerdings konnte dieses Ergebnis nicht für alle Stimuli, die nach der zweiten Methode bearbeitet worden sind, erzielt werden.

4 Diskussion

Bei dieser Pilotstudie handelt es sich zwar nur um eine erste und grobe Annäherung an die akustischen Ausprägungen prototypischer emotionaler Vokalisierungen wie sie von Schauspielern produziert wurden. Dennoch kann eine solch explorative Vorgehensweise mit dazu beitragen zu verstehen wie wir affektive Äußerungen verarbeiten. Die rhythmische Struktur scheint hierbei eine zentrale, aber wohl nicht die einzige Rolle zu spielen. Weitere Perzeptionsexperimente, wie sie beispielsweise in [5] für Lachen durchgeführt wurden, sind notwendig, um die Verarbeitung und die Verwechselbarkeit der beiden in der Valenzbewertung so unterschiedlichen Affektäußerungen zu verstehen.

In den protypischen Beispielen der Montreal Affective Voice Database [3] ist die Emotionskategorie "glücklich" von allen zehn Schauspielern mit Lachen dargestellt worden. An dieser Stelle sollte angemerkt werden, dass Lachen als non-verbale Schauspieläußerung außer für Freude auch noch für andere affektive Zustände wie Erfolg, Schadenfreude, Kitzeln, Spott wie benutzt wird. Entsprechende Untersuchungen stellen Unterschiede sowohl in ihrer akustischen Ausprägung [9] als auch in ihrer perzeptiven Zuordnung zu Fotos mit emotionsgeladenen Gesichtern [6,7].

Noch relativ unerforscht ist die Wirkung von Affektäußerungen in synthetischer Sprache. Es scheint sich zwar allmählich die Erkenntnis zu etablieren, dass zu einer Anpassung synthetischer Sprache in Dialogsystemen an menschliche Sprachproduktion auch non-verbale Vokalisierungen gehören [4] – weitgehend unklar bleibt jedoch die Frage wie diese zu modellieren sind. Dabei scheinen vor allem Lachen und Weinen für die Memorierbarkeit akustisch übermittelter sprachlicher Information eine sehr günstige Rolle zu spielen [1].

Literatur

[1] Armony, J.L., Chochol, C., Fecteau, S. & Belin, P. (2007). Laugh (or Cry) and You Will Be Remembered: Influence of Emotional Expression on Memory for Vocalizations. Psychological Science 18(12), pp. 1027-1029.

[2] Bachorowski, J.-A., Smoski, M.J. & Owren, M.J. (2001). The acoustic features of human laughter. Journal of the Acoustical Society of America 111(3), pp. 1582-1597.

[3] Belin P., Fillion-Bilodeau S. & Gosselin F. (2008) The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods 40(2), pp. 531-539.

[4] Campbell, N. (2006). Conversational speech synthesis and the need for some laughter. IEEE Transactions on Audio, Speech, and Language Processing 14(4), pp. 1171-1178.

[5] Kipper, S. & Todt, D. (2003). The role of rhythm and pitch in the evaluation of human laughter. Journal of Nonverbal Behaviour 27, 255-272.

[6] Sauter, D. (2010). More than happy: The need for disentangling positive emotions. Current Directions in Psychological Science 19, 36-40.
[7] Sauter, D., Eisner, F., Ekman, P. & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. Proceedings of the National Academy of Sciences, 107(6), 2408-2412.
[8] Schuller, B., Batliner, A., Steidl, S. & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication (http://dx.doi.org/10.1016/j.specom.2011.01.011).
[9] Szameitat, D.P., Alter, K., Szameitat, A.J., Wildgruber, D., Sterr, A. & Darwin, C.J. (2009). Acoustic profiles of distinct emotional expressions in laughter. Journal of the Acoustical Society of America 126 (1), pp. 354-366.
[10] Trouvain, J. (2003). Segmenting phonetic units in laughter. Proc. 15th International Congress of Phonetic Sciences(ICPhS), Barcelona, pp. 2793-2796.