Amplitude and Amplitude Variation of Emotional Speech

Hartmut R. Pfitzinger[1] & *Christian Kaernbach*[2]

[1]Institute of Phonetics and Digital Speech Processing (IPDS)
[2]Institute of Psychology
Christian-Albrechts-University at Kiel, Germany
hpt@ipds.uni-kiel.de, www.kaernbach.de

Abstract

The present study introduces a recording technique to maintain all dynamic information of even full-blown emotional speech, and investigates the effect of emotion class, speaker, and sentence type on the amplitude of speech. The results show that the factors emotion class and speaker are highly significant and that the former explains half of the variance while the latter explains only one ninth. Amplitude, as simple as it is, should therefore not be neglected or normalized in acoustic recordings and analyses of emotional speech.

Index Terms: emotional speech, affect, acoustic features, amplitude, intensity, speech recording

1. Introduction

Emotions add up to ca. 30 dB amplitude variation to the typical speech amplitude range. Previous studies on emotional speech were not able to preserve these dynamics during the recording process: either loud speech passages were clipped, or soft passages were masked by the noise floor of the recording equipment, or both. To reduce these problems, recording gain is usually readjusted for each speaker and emotion, or dynamic range is compressed, both resulting in loss of amplitude information.

In addition, speech amplitude (or intensity) as a prosodic contour is supposed to be an acoustic property of subordinate communicative importance compared with pitch contours or local speech rate variation [16]. This becomes immediately evident when realizing that broadcasting has no noticeable impact on the meaning or communicative function of speech despite the fact that the short-term amplitude is generally strongly manipulated via dynamic range compression algorithms to maximize loudness of the transmitted signal. Accordingly, Grimm et al. 2007 [11] who used emotional utterances taken from TV shows with strong dynamic compression, found little impact of amplitude related features on automatic emotion recognition. However, emotions are accompanied by amplitude variation as shown by Cahn 1990 [6] and Burkhardt 2005 [4] with synthetic speech. Hammerschmidt & Jürgens 2007 [12] and many others found a highly significant effect of amplitude on various emotions. Thus, the present investigation addresses amplitude and amplitude variation of emotional speech.

1.1. Basic emotions, dimensions, appraisal processes

According to Scherer et al. [21, 22, 23] facts and circumstances are subjected to appraisal processes which determine the emotions people feel. A specific set of appraisal criteria is also said to be important in emotion differentiation.

Dimensional approaches describe emotions using two or three continuous-valued emotion primitives, namely *valence* (negative↔positive), *activation* (high↔low), and sometimes *dominance* (strong↔weak). Various dimensional systems with partly synonymic but also heteronymic dimensions exist, and it is evident that some dimensions are unintuitively and arbitrary.

Plutchik developed his theory of basic emotions from 1958 to 1994 [18, 19] within the framework of evolutionary psychology. Although it is based on eight *primary* or *basic* emotions, he introduced the intensity of emotion as a continuous dimension which infinitely increases the number of possible emotions.

2. An emotional speech database

A new emotional speech database became necessary, as an inspection of existing databases revealed that they were either too small or suffered from speech amplitude compression.

We decided to record acted emotions motivated by the facts that 1) speech material is easy to control for comparative studies, and 2) high-quality recordings are possible in a studio environment. The shortcomings that acted emotions are not authentic and possibly hyper-articulated are obvious. And as long as no perceptual or psychophysiological data (heart rate variability, electrodermal activity, ...) are available, analyses are based on *intended*, not necessarily *realized* emotional expressions.

The choice of our set of basic emotions was inspired by the relevant literature on speech and emotion [1, 2, 3, 5, 8, 9, 15, 24] which revealed the four basic emotions *anger, fear, joy,* and *sadness* as a common denominator. These are a subset of Plutchik's 1958 [18] basic emotions which additionally comprise *disgust, surprise, curiosity/anticipation,* and *acceptance/trust*. While investigations on speech and emotion hardly ever consider the latter two categories, *disgust* and *surprise* are often attempted to be included. We omitted *surprise* as in its positive variant it is close to joy, and in its negative variant close to sadness.

Fig. 1 outlines all independent variables controlled during the recordings of the emotional speech database. A unique feature is that short dialogues have been acted by two speakers.

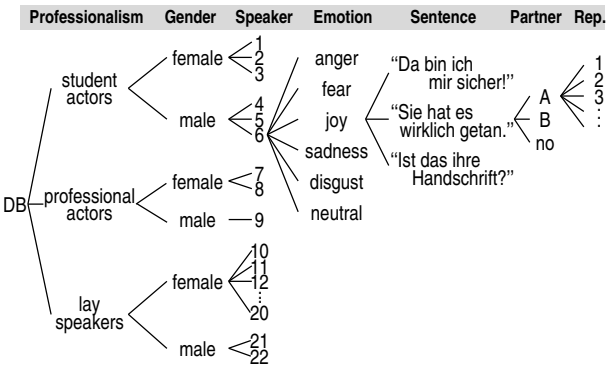

Figure 1: *Independent variables and the extent of the database.*

3. Technical design aspects

High-quality hybrid recording devices which combine microphone amplifiers with 24-bit-A/D-converters (e.g. *RME Fireface 800*) at best reach an A-weighted dynamic range between 112 and 117 dBA. But the total harmonic distortion plus noise (THD+N) typically ranges from 104 to 108 dB which in fact determines the dynamic range effectively available for measurements of speech amplitude.

The sound pressure level (dB SPL) is a logarithmic measure relative to a well-defined reference sound pressure of $p_0 = 20\ \mu$Pa which corresponds to a just audible sinusoidal oscillation at 2 kHz.

The *Microtech Gefell M 940* is a high-quality large-membrane condenser microphone with cardioid directional characteristics and ultra low noise. It provides an SNR (signal-to-noise ratio) of 88 dBA relative to a 1 kHz sinusoidal oscillation with a sound pressure of 1 Pa. Since 1 Pa is 94 dB louder than 20 μPa it corresponds to 94 dB SPL. Thus, the microphone produces equivalent noise of 6 dBA SPL (cp. *Røde NT1-A*: 5 dBA SPL, *Neumann TLM 103*: 7 dBA SPL). At 3.4 kHz the ear is maximally sensitive and perceives signals with a level of -6 dB SPL (the thermal noise caused by Brownian molecular motion is only 12 dB softer). As a consequence, the best microphones available are 11 dB to 13 dB worse than the human ear.

Although the threshold of pain is 100 Pa (134 dB SPL) which is 5.000.000 times more than 20 μPa, the recording hardware does not have to cope with these extreme levels since even shouting hardly exceeds 110 dB SPL. Consequently, when recording shouted speech the maximum SNR of today's recording equipment reaches 103 dB (110 dB minus 7 dB). For this specific application the main limitation of the dynamic range is the noise-floor of the used microphone.

3.1. Mouth-microphone distance

The sound pressure level of normal talking ranges from roughly 40 to 60 dB SPL at 1 m measuring distance. Since halving the mouth-microphone distance increases the level by 6 dB, a distance of 25 cm yields a typical sound pressure of 52 to 72 dB SPL. Thus, even the best recording equipment in a perfectly quiet vocal booth yields an SNR of only 47 dB which corresponds to only 8 bit amplitude resolution.

However, reducing the mouth-microphone distance increases the amplitude variation caused by head movements, i.e. $\pm$5 cm at a distance of 1 m modulates the recorded amplitude between +0.44 dB and -0.42 dB while $\pm$5 cm at 25 cm results in +1.93 dB to -1.58 dB, the latter being unacceptable. Therefore, we chose a 50 cm distance representing a virtual optimum between noise-floor reduction (by 6 dB in comparison to a 1 m distance) and the influence of head movements ($\pm$5 cm then correspond to +0.91/-0.83 dB).

3.2. Two-microphone recordings

Each speaker is recorded with two microphones as shown in Fig. 2 which are almost at the same position but have very differently adjusted gain factors. The purpose of this procedure is that the more sensitive microphone guarantees the highest possible signal-to-noise ratio at soft speech passages while the insensitive microphone records even the loudest speech without any clipping artifacts. The gain difference was manually set to ca. 24 dB. Once set, the gain adjustment was left unchanged during all recording sessions and the mean mouth-microphone distance was kept constant at 50 cm with a typical head movement variation during each session of about $\pm$5 cm.

3.3. Delay estimation

The delay between the two channels was estimated via a normalized cross-correlation. Since local variation of the delay is possible due to up-down head movements a local delay estimation was applied (Pfitzinger & Reichel 2006 [17]). Every 10 ms, stretches of the two speech signals with 200 ms duration were submitted to coss-correlation. The delay turned out to be almost constant and between 0 and 1 sample (at a 48 kHz sampling rate) almost regardless of the speaker. One sample corresponds to a mouth-microphone distance difference between the two microphones of 7.08 mm (340 m/s divided by 48000 cycles/s). But since the distance between the centers of the two microphone membranes was 35 mm, only an up-down head movement of $\pm$5.1 cm at a mouth-microphone distance of 50 cm would cause this variation. Typically, vertical head movements are smaller. Furthermore, the delay variation is even too small to detect any significant effect of the different heights of the speakers.

3.4. Amplitude processing

In order to estimate the precise amplitude difference between the two channels linear regression could not be applied directly to the speech signal samples since microphone tolerances even of pairwise factory-selected microphones are too large to achieve identical signals. The estimation of the scaling factor via direct linear regression is distorted by this and by the very small delay of 0 to 1 samples.

Therefore, another method was used: both speech signals were full-wave rectified and smoothed using a 20 ms Kaiser window ($\beta = 5$). Then they were submitted to first-order linear regression whose 2nd coefficient is the scaling factor if 1) linear instead of logarithmic amplitude values are used (the former differ by a factor while the later differ by an addend), 2) clipped stretches are omitted (they disturb the estimation of the scaling factor severely), and 3) stretches with very low energy are skipped (since the resulting amplitude values are unreliable).

The final signals were estimated from the 24-bit recordings by first converting them to numerical floating-point format, then applying linear-phase FIR high-pass filtering (-3 dB at 79 Hz) to remove subsonic noise, then performing global amplitude scaling to reduce the headroom between the highest amplitude of all signals and the maximum of the digital fixed-point format to 0.1 dB, and finally dithering to 16 bit with a triangular probability density function to increase dynamic range and avoid any distortion or musical noise. The gain of the more sensitive channel was adjusted to be 24 dB louder providing the max. SNR.

Figure 2: *Recording of emotional speech: two microphones are directed towards each speaker.*

4. Acoustic analysis

The six student actors were selected for acoustic analysis. Their recordings were cut into 1837 single utterances by the following two steps: In the first pass, all signals were automatically pre-segmented applying Schmitt triggering with 5 dB hysteresis to the short-term amplitude smoothed over 350 ms. In the second pass, false alarms were manually deleted and the resulting utterance boundaries were re-adjusted as well as labelled.

4.1. Amplitude measurements

In order to estimate a single value representing the amplitude of an utterance we compared three different methods. Root mean square amplitude, mean absolute amplitude, and amplitude envelope were extracted framewise from each utterance in steps of 10 ms. Since frames coming from the short speech pauses of ca. 50 ms at the beginning and end of each utterance, or from plosive closures, have very low amplitude values and would distort the mean amplitude of all frames, only 20% of the loudest frames were subjected to averaging, representing the mean of the loudest 20% of each utterance. Actually, this proportion can be chosen within a wide range of 10% to 40% as it has little or no impact on the amplitude ratios between different utterances. Moreover, it turned out that the three amplitude measurement methods yield highly correlated values. Thus, the actual choice has negligible impact on subsequent statistical analysis.

4.2. Results

The 1837 amplitude values were submitted to statistical analysis. A one-way ANOVA based on GLM was applied with the independent factors *emotion*, *speaker*, and *sentence*. As expected the factors *emotion* and *speaker* as well as their interaction are highly significant ($p<0.001$) as shown in Tab. 1. The factor *sentence* is statistically not significant and thus not shown in Fig. 3.

The interaction between the factors *speaker* and *emotion* is statistically highly significant and explains 7% of the total variance. This means that the different speakers use amplitude in a different way to express emotions.

This finding is also reflected in Fig. 3: The mean amplitude values of all neutral utterances differ by only ± 2 dB across speakers. However, in emotional speech the speaker differences range from ± 5 to ± 7 dB. Consequently, a normalization procedure which adjusts speaker-individual gain factors to equal amplitude for the neutral utterances could not significantly reduce the unexplained variance, which is 31%.

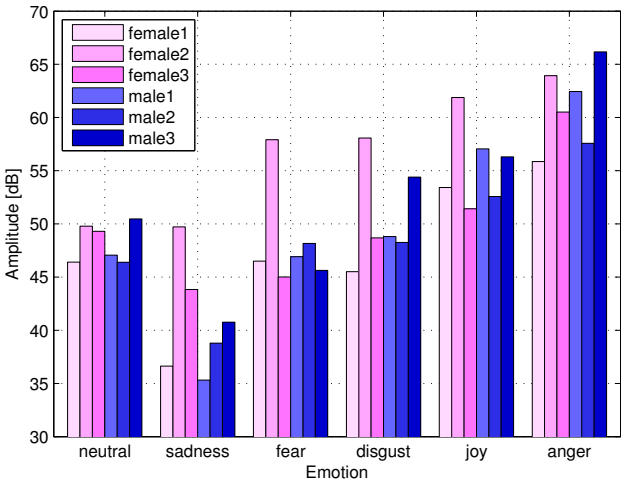

Figure 3: *Average speech signal amplitudes of 6 student actors producing 658 neutral and 1179 emotional utterances.*

	df	F	p	Variance explained
emotion	5	750.187	0.000	47.71%
speaker	5	172.937	0.000	11.00%
sentence	2	2.808	0.061	0.07%
emotion×speaker	25	22.479	0.000	7.15%

Table 1: *One-way ANOVA applied to amplitude values of 1837 utterances. The three remaining interactions together explain only 3.11% of the total variance and are omitted.*

5. Discussion

One of the notable results is that student actors use amplitude in different ways to express emotions. For instance, speaker *female2* (see Fig. 3) uses a remarkably high amplitude to express *sadness, fear, disgust,* and *joy* compared with the other five speakers, but her amplitude for *neutral* and *anger* is in line with the other speakers.

The most obvious explanation is individual personality: e.g. speaker *male2* was reported to hardly ever showing anger which could be the reason why the mean speech amplitude of his angry utterances was roughly 5 dB lower than that of speaker *male1* and more than 8 dB lower than that of speaker *male3*.

Another reason could be that it is implicitly quite clearly defined to any speaker how to produce neutral utterances, but a student actor might have several, very different, strategies to choose from when expressing a specific emotion.

A broader interpretation of the amplitude values within the framework of dimensional emotion classification is unprofitable since amplitude is not directly correlated with any of the coordinates of any dimensional emotion system. Typically, anger and fear are both judged to be almost equally active and negative, but they show very different amplitude values (see Fig. 3), while fear is negative on the valence dimension but has no statistically significant amplitude difference compared with neutral utterances (p=0.85, Scheffé's test of all factor level contrasts).

Amazingly, our results give first indication that Plutchik's three-dimensional arrangement of the basic emotions [19] is in approximate accordance with amplitude if the amplitude dimension is oriented diagonally from sadness to anger in his scheme. Then, disgust is half way between them and fear is close to sadness while joy is close to anger (compare with Fig. 3).

While, in informal interviews, some actors admitted severe difficulties to express disgust, lay speakers mentioned other emotions as being more difficult to express. This might be due to the fact that actors are aware of the different nature of disgust in contrast to anger, joy, fear, and sadness: typically, it is a short-term affective disturbance which Scherer calls *affect burst*. Banse & Scherer 1996 [1] achieved extraordinarily low perceptual recognition of disgust in long sentences (15%) which they attribute to this fact.

6. Conclusions

The present acoustic analysis examined mean speech amplitude and its variation as caused by different intended emotions of six student actors. For statistical analysis, a neutral speaking style and the five basic emotions *anger, fear, joy, sadness,* and *disgust* were chosen as levels of the independent factor *emotion* while three male and three female student actors constituted another

independent factor *speaker*. Both factors appear to have highly significant influence on speech amplitude. But while *speaker* explains only one ninth of the total variance, *emotion* accounts for half of the variance.

This implies that, first, acted speech is accompanied by a strong correlation between intended expression and amplitude. Second, the used method for measuring speech amplitude is obviously appropriate to condense the local short-term amplitude variations of utterances into one single representative value. The unexplained variance of 31% suggests improving and evaluating this method further and also to test other independent factors such as *gender*, although the present emotion database is not large enough to reliably draw conclusions on this factor.

If the expression of emotion and affect is not only a reflection of the inner state but also a listener-directed information then it can be assumed that in natural communicative situations the emotional speaker adapts features to the communicative environment (H&H theory [14]). But the interaction between the type of emotion and the communicative environment still remains an open research issue.

Concerning the recording procedure of emotional speech in studio environment it is of major importance to use a perfectly quiet vocal booth or anechoic chamber and an ultra low-noise microphone. The choice of the soundcard is not crucial as long as it reaches a THD+N of at least 104 dB because then, noise floor is not the critical limitation even if gain is set to 36 dB headroom (6 bit) above normal speech amplitude to avoid any clipping during the loudest passages of maximally active emotions, as the remaining 11–12 bit guarantee a final SNR of at least 68 dB at amplitude peaks of normal speech. Our two-microphone technique adds approximately 6 dB reaching 74 dB which results in an SNR of 50 dB for the softest utterances arising from the emotion category *sadness*.

7. Outlook

3580 utterances of all 22 speakers will be available soon. It enables another independent variable *speaker professionalism* to be statistically tested. During the follow-up experiments all utterances are subject to perceptual and psychophysiological analysis. The resulting subjective and objective measures allow analysis of the interaction between intended, perceived, and felt emotions. The outcomes guide further acoustic analyses of the speech signals to achieve more insight in the relations between acoustic features and emotional content of speech. Concretely, the recorded speech data is suited to estimate the correlation between speech amplitude and voice quality, particularly normalized amplitude quotient, breathyness, and spectral slope [7, 10].

8. Acknowledgements

Thanks to Theresa Anders, Julia Bergherr, Jessica Bösel, Susanne Huth, Thomas Jacobsen, Franziska Jähne, Sophie Lütt, and Nina Redlingshöfer from IPDS, Kiel for manually segmenting and labelling the emotional speech database on a sentence level.

9. References

[1] Banse, R.; Scherer, K. R. 1996. Acoustic profiles in vocal emotion expression. *J. of Personality and Social Psychology*, 70(3): 614–636.

[2] Bänziger, T.; Scherer, K. R. 2005. The role of intonation in emotional expressions. *Speech Communication*, 46: 252–267.

[3] Boula de Mareüil, P.; Célérier, P.; Toen, J. 2002. Generation of emotions by a morphing technique in English, French and Spanish. In *Proc. of the 1st Int. Conf. on Speech Prosody*, pp. 187–190, Aix-en-Provence.

[4] Burkhardt, F. 2005. Emofilt: the simulation of emotional speech by prosody-transformation. In *Proc. of Interspeech '05*, pp. 509–512, Lisbon; Portugal.

[5] Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; Weiss, B. 2005. A database of German emotional speech. In *Proc. of Interspeech '05*, pp. 1517–1520, Lisbon; Portugal.

[6] Cahn, J. E. 1990. Generating expression in synthesized speech. Master's thesis, Massachusetts Institute of Technology, Cambridge; Massachusetts.

[7] Campbell, N.; Mokhtari, P. 2003. Voice quality: the 4th prosodic dimension. In *Proc. of the XVth Int. Congress of Phonetic Sciences*, vol. 3, pp. 2417–2420, Barcelona.

[8] Cowie, R.; Cornelius, R. R. 2003. Describing the emotional states that are expressed in speech. *Speech Communication*, 40: 5–32.

[9] Erickson, D. 2005. Expressive speech: Production, perception and application to speech synthesis. *Acoustical Science and Technology*, 26(4): 317–325.

[10] Gobl, C.; Ní Chasaide, A. 2003. The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40(1–2): 189–212.

[11] Grimm, M.; Kroschel, K.; Mower, E.; Narayanan, S. 2007. Primitives-based evaluation and estimation of emotions in speech. *Speech Communication*, 49(10–11): 787–800.

[12] Hammerschmidt, K.; Jürgens, U. 2007. Acoustical correlates of affective prosody. *J. of Voice*, 21(5): 531–540.

[13] LeDoux, J. E. 1996. *The emotional brain*. Simon and Schuster, New York.

[14] Lindblom, B. E. F. 1990. Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle, W. J.; Marchal, A., eds., *Speech production and speech modelling*, Nr. 55 in Nato ASI series D: Behavioural and social sciences, pp. 403–439. Kluwer Academic Publishers, Dordrecht, Boston, London.

[15] Mozziconacci, S. J. L. 2002. Prosody and emotions. In *Proc. of the 1st Int. Conf. on Speech Prosody*, pp. 1–9, Aix-en-Provence.

[16] Pfitzinger, H. R. 2006. Five dimensions of prosody: Intensity, intonation, timing, voice quality, and degree of reduction. In Hoffmann, R.; Mixdorff, H., eds., *Speech Prosody Abstract Book. Studientexte zur Sprachkommunikation*, vol. 40, pp. 6–9. TUDpress, Dresden.

[17] Pfitzinger, H. R.; Reichel, U. D. 2006. Delay compensation between the speech signal and the corresponding electroglottograph signal. In *Proc. of the AST Workshop*, pp. 63–74, Maribor; Slovenia.

[18] Plutchik, R. 1958. Outline of a new theory of emotion. *Trans. of the New York Academy of Scienes*, 20: 394–403.

[19] Plutchik, R. 1994. *The psychology and biology of emotion*. Harper Collins College Publishers, New York.

[20] Rossato, S.; Audibert, N.; Aubergé, V. 2004. Emotional voice measurement: A comparison of articulatory-EGG and acoustic-amplitude parameters. In *Proc. of the 2nd Int. Conf. on Speech Prosody*, pp. 749–752, Nara; Japan.

[21] Scherer, K. R. 2003. Vocal communication of emotion: a review of research paradigms. *Speech Communication*, 40(1–3): 227–256.

[22] Scherer, K. R.; Johnstone, T.; Klasmeyer, G. 2003. Vocal expression in emotion. In Davidson, R. J.; Scherer, K. R.; Goldsmith, H. H., eds., *Handbook of affective sciences*, chap. 23, pp. 433–456. Oxford University Press, Oxford, New York.

[23] Scherer, K. R.; Schorr, A.; Johnstone, T., eds. 2001. *Appraisal processes in emotion. Theory, methods, research*. Oxford University Press, Oxford, New York.

[24] Schröder, M. 2004. *Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis*. Ph.D. thesis, Institut für Phonetik, Univ. des Saarlandes.