OPEN ACCESS
OPUS AUGSBURG

Universität Augsburg
Universitätsbibliothek

The perception of emotions in noisified nonsense speech

Emilia Parada-Cabaleiro, Alice Baird, Anton Batliner, Nicholas Cummins, Simone Hantke, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Parada-Cabaleiro, Emilia, Alice Baird, Anton Batliner, Nicholas Cummins, Simone Hantke, and Björn Schuller. 2017. "The perception of emotions in noisified nonsense speech." In *Interspeech 2017, 20-24 August 2017, Stockholm*, edited by Francisco Lacerda, 3246–50. ISCA. https://doi.org/10.21437/interspeech.2017-104.

Nutzungsbedingungen / Terms of use:

licgercopyright

Dieses Dokument wird unter folgenden Bedingungen zur Verfügung gestellt: / This document is made available under these conditions:
Deutsches Urheberrecht
Weitere Informationen finden Sie unter: / For more information see:
https://www.uni-augsburg.de/de/organisation/bibliothek/publizieren-zitieren-archivieren/publiz/

INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

The Perception of Emotions in Noisified Nonsense Speech

Emilia Parada-Cabaleiro[1], *Alice Baird*[1], *Anton Batliner*[1],
Nicholas Cummins[1], *Simone Hantke*[1,2], *Björn W. Schuller*[1,3]

[1]Chair of Complex and Intelligent Systems, University of Passau, Germany
[2]Machine Intelligence & Signal Processing Group, Technische Universität München, Germany
[3]Machine Learning Group, Imperial College London, U.K.

Emilia.ParadaCabaleiro@uni-passau.de

Abstract

Noise pollution is part of our daily life, affecting millions of people, particularly those living in urban environments. Noise alters our perception and decreases our ability to understand others. Considering this, speech perception in background noise has been extensively studied, showing that especially white noise can damage listener perception. However, the perception of emotions in noisified speech has not been explored with as much depth. In the present study, we use artificial background noise conditions, by applying noise to a subset of the GEMEP corpus (emotions expressed in nonsense speech). Noises were at varying intensities and 'colours'; white, pink, and brownian. The categorical and dimensional perceptual test was completed by 26 listeners. The results indicate that background noise conditions influence the perception of emotion in speech – pink noise most, brownian least. Worsened perception invokes higher confusion, especially with sadness, an emotion with less pronounced prosodic characteristics. Yet, all this does not lead to a break-down of the 'cognitive-emotional space' in a Non-metric MultiDimensional Scaling representation. The gender of speakers and the cultural background of listeners do not seem to play a role.

Index Terms: emotion in nonsense speech, background noise, signal masking, perception test.

1. Introduction

The ability to perceive different emotions expressed by others based upon non-linguistic information, such as facial expressions and vocal (non-verbal) cues, is a natural process for humans. The perception and identification of emotion in others allows for the occurrence of empathetic connections that link perceived emotion and personal knowledge. This skill is essential for having effective human-human interactions, and people lacking this ability often have difficulties building social relationships.

Auditory impairments affect millions of people worldwide, an estimated 360 million individuals in 2011 [1]. Such impairments are often linked to a reduced ability to perceive emotional expression in others [2]. For individuals with no hearing disabilities, noise polluted environments can create difficulties in perceiving emotions, which could be compared to those with hearing-impairment. Unfortunately, noise pollution is an indiscernible part of our daily life, caused often by human interventions such as traffic, industrial work, leisure activities, and other background noise.

It is also well known that background noise can cause damage to human well-being, both physically and psychologically [3]. In these situations, the ability to perceive the linguistic and emotional message of vocal and verbal content is reduced. However, the perception of emotional speech in adverse environmental conditions is still an almost completely unexplored field. Despite a variety of studies which have been made to evaluate the background noise effect in linguistic understanding [4, 5, 6, 7], it seems that no similar research has been made for the perception of emotional speech.

The effect of *white*, *pink*, and *brownian* noise in perception has been extensively studied in different areas, such as sound-therapy [8] or the linguistic understanding in background noise [4, 5, 6, 7]. Nevertheless, to the best of our knowledge, artificially created background noise environments have not been considered in the evaluation of emotion perception in speech. Several studies have been conducted, altering the original signal through a variety of electronic techniques [9, 10, 11], such as *random splicing* [12] or *reversing* [13] among others. However, none of these studies considers the addition of artificial noise in the evaluation of listener perception of emotions in speech. The evaluation of emotional speech in background noise has previously only been considered in the realm of automatic speech emotion recognition, in which the presence of noise is a well known confounding factor [14, 15, 16].

Motivated by the lack of research involving speech-based emotion perception in adverse environmental conditions, in the presented study we evaluate emotion perception in the presence of artificial background noise (white, pink, and brownian), employing a categorical/dimensional forced-choice test. Section 2 covers emotion models, database, and synthetic manipulations. In section 3, we present the perception study; section 4 discusses the results. Finally, section 5 offers a succinct conclusion and outline our future work plans.

2. Methodology

2.1. Factors influencing the perception of emotional speech

In designing our listening test, we take into account the two main emotion models: the *categorical model* and the *dimensional model* [18]. The categorical model discriminates between primary (basic) and secondary emotions [19]. Basic emotions are considered being universal, i. e., common for all cultures; there is a general agreement for anger, sadness, happiness, and fear being basic emotions [20, 19]. Secondary emotions (like jealousy) are more complex, and comprised of a combination of basic emotions [21]. Emotions such as disgust and surprise are contentious, being considered by some authors as basic [19], by others as secondary [22]. The dimensional model, on the other hand, places emotions in a continuous hyperplane characterised by different 'dimensions' [23]. The most common dimensional model is the bi-dimensional model [23], with the first dimension being arousal (related to the intensity of an

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-104

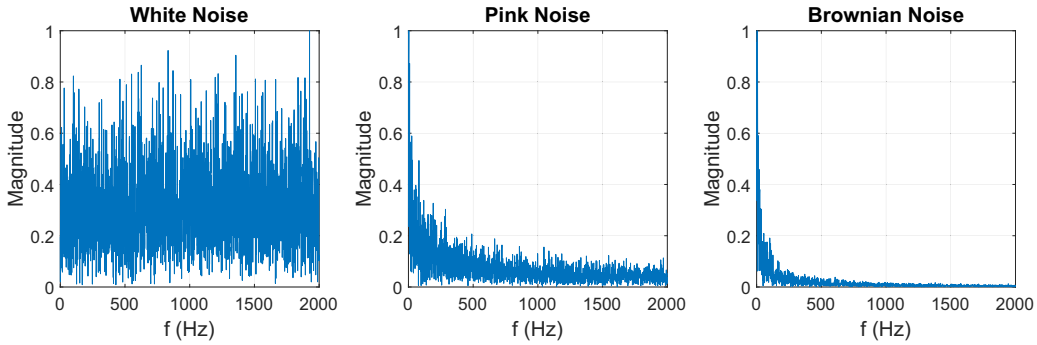

Figure 1: *A comparison of the spectral distribution between 0–2 kHz for the white, pink, and brownian noise types.*

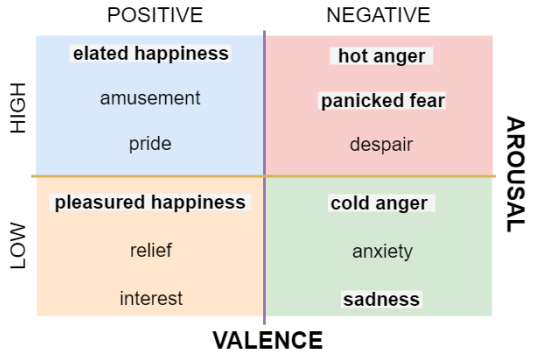

Figure 2: *Correspondence between emotion categories and bi-dimensional model of the 12 emotional states considered in GEMEP [17]. Six of them (highlighted) have been used in our listening test: three for each level of arousal (high and low); two and four for positive and negative valence, respectively.*

emotional state), and the second dimension valence (related to hedonistic values). For our study, we considered primary and secondary emotions, as well as the dimensions of arousal and valence.

The emotion theory taken into consideration will influence the strategy used to evaluate the perception of listeners. The forced-choice test is typically used in tests involving the categorical model [24]. However, this type of test has been criticised for studying *discrimination* instead of *recognition* [24]. In order to avoid the strict one-to-one relationship between the forced-choice test and the emotional labels, some studies offer the listener a possibility to choose the label 'neutral'. Nevertheless, this label could be used by the listener as a strategy to avoid any decision between emotion categories – thus it could mean 'undecided' as well as 'neutral' [25]. Another strategy to avoid influencing the listeners by leading or restricting their choice is to include so-called *distractor labels* that have not been considered previously during data set configuration [26].

Finally, listener perception of emotions is also highly related to culture. There is evidence that the level of agreement between listeners from different cultures is higher than chance [27], especially in similar cultures [28]; the effects of cultural differences are discussed in [29, 30]. However, other studies show that perception of emotions by native speakers is much more precise than by non-native speakers [31]. A solution to overcoming this issue is to consider nonsense utterances, which avoid linguistic influence in perception, since the data is not linked to any specific language and does not contain any contextual meaning [27].

Table 1: *Percentage of accuracy in the recognition of the considered emotions: CO (cold anger), EL (elated happiness), HO (hot anger), PA (panicked fear), PL (pleasurable happiness), and SA (sadness); in both clean (cl) and -1 dB SNR background noise conditions (br-brownian, pi-pink, wh-white) are displayed.*

%	CO	EL	HO	PA	PL	SA	mean
cl	64.3	30.1	59.5	29.8	28.1	58.6	45.1
br	61.3	18.5	34.1	21.7	18.9	66.2	36.8
pi	40.1	11.2	23.1	19.3	15.1	59.0	28.0
wh	51.0	17.3	31.5	20.0	12.6	63.6	32.7

2.2. GEMEP database

The *GEneva Multimodal Emotion Portrayals* (GEMEP) database [17] (used in the `ComParE` 2013 challenge [32]) has been chosen for our research as it takes into consideration both emotional models and consists of nonsense utterances to avoid any cultural bias. GEMEP is an acted database, including emotional utterances pronounced by five male and five female professional native-French speaking actors. Both categorical and dimensional models are considered in this database. Six of the eight speakers, three males (GEMEP ID: 01, 03, 04) and three females (GEMEP ID: 02, 06, 10) have been randomly chosen, to restrict effort and time needed for experimental sessions. The nonsense utterance *ne kal ibam soud molen!* was chosen for our experiments.

2.3. Synthetic manipulation

The utterances have been masked by three different types of noises as displayed in Figure 1: *white noise* presents a flat spectrum containing all the frequencies at the same level (20 Hz - 20 kHz); *pink noise* presents a spectrum level with negative slope of 10 dB per octave (1/f noise); and *brownian noise* is characterized by higher energy in the lowest frequencies having each octave as much energy as the two octaves above, that is the energy falls at 6 dB per octave (1/f2 noise). These noises, unlike others like blue, violet or grey, favour the lower frequencies that are most relevant for speech; thus, they have been chosen.

Four different Signal-to-noise ratio (SNR) levels (-1 dB, -0.5 dB, +1 dB, and +3 dB) have been applied to each of the three noises. These levels were selected in a pre-evaluation stage considering the maximum and minimum levels as the limit threshold to perform a meaningful perception test, beyond which no differences are perceived. Two intermediate SNR (-0.5 dB and +1 dB) have been introduced as transitional levels between the maximum and minimum previously selected. The artificial background noisy conditions have been created in

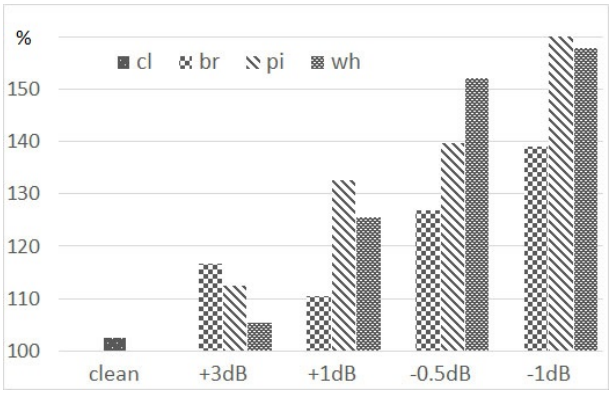

Figure 3: *Percentage of emotions identified as sadness for each level and type of noise (br-brownian, pi-pink, wh-white) and for the clean (cl) condition (36 utterances encoded in each bar).*

Table 2: *Sums of 'identified as' (hits and false alarms) in percentage for the six emotions, and for clean and the three noise conditions at -1 dB SNR (36 cases represented in each row).*

%	CO	EL	HO	PA	PL	SA
cl	124.6	34.2	78.0	86.0	83.8	102.6
br	147.4	29.2	46.5	77.9	78.6	139.0
pi	135.6	23.8	37.2	75.7	68.1	168.9
wh	126.9	27.1	49.3	80.9	65.0	157.8

Matlab R2014a [33]. Samples of each noise type were normalized to -1 dB and noise was applied to each clean sample at the specified SNR value.

Considering 6 emotional utterances (one for each emotional state), 6 actors, 3 different noises (white, pink, and brownian), 4 applied SNR (-1 dB, -0.5 dB, +1 dB, and +3 dB), plus 36 clean utterances (18 produced by males and 18 by females), each listening session is made up of 468 stimuli.

3. Perception study

To ensure a consistent listening environment, the tests were performed in an acoustically treated recording studio with identical conditions and equipment such as closed (noise isolating) earphones and identical computer set-up for all participants. The volume was presented at a comfortable unchanged level and the test was performed on a browser based interface provided through our gamified crowdsourcing platform *iHEARu-PLAY* [34]. The listeners had the possibility of repeating each utterance and the annotation indefinitely; the samples were presented randomly. In total 26 subjects took part, 13 from Germany and 13 from other nationalities[1]. Ages were between 22 and 31 years, with a mean age of 23.7 years and a standard deviation of 2.5 years. The participants, who voluntarily gave up their time, were students or employees from the University of Passau who had no insight into the research procedure. In order not to overload the listeners, the sessions were designed to last no more than one hour; participants were also able to take a break whenever needed.

The listening test itself was a forced-choice task; ten emotion categories have been considered in order to reduce the likelihood that the judgement is due to discrimination instead of recognition [24]. Six of the ten categories – *cold anger* (CO), *elated happiness* (EL), *hot anger* (HO), *panicked fear* (PA), *pleasurable happiness* (PL), and *sadness* (SA) – are classes from the emotions considered in GEMEP (cf. Figure 2). The other four – *desperate sadness*, *worried fear*, *surprise*, and *disgust* – are distractor classes. Surprise and disgust have been considered as distractors because their ambiguity could generate an interesting challenge to the listeners. Worried fear and desperate sadness have been added as distractors in order to complete the missing arousal levels for the basic emotions sadness and fear, creating in this way a balanced forced-choice test.

[1] 5 from India, 2 from Tunisia, 2 from Spain, 1 from Iran, 1 from Mexico, 1 from Russia, and 1 from UK.

4. Results and discussion

Our analysis reveals that noisy environments influence listeners' perception of emotion (cf. Table 1). In all noise type conditions, all the emotions evaluated except sadness are less easily perceived. Indeed, sadness displays a better performance in noisy conditions than in non-noisy environments. This could be due to sadness – characterized by low tone, pitch, and energy – being perceived as an attenuated emotion, which makes background noise less influential than it is the case for other, higher aroused emotions. In fact, at higher SNR, the percentage of utterances wrongly identified as sadness increases (cf. Figure 3).

Furthermore, another low aroused emotion, cold anger, shows by far higher levels of confusion with respect to the other emotional categories: in Table 2 it is shown that the sum of the emotions identified as cold anger and sadness is greater than 100% and increases in noisy environments whereas for the rest of the emotions, this figure is always below 100% and decreases in noisy environments for all the noises considered. The distractors were not chosen often, neither in clean condition nor in strong background noise (-1 dB SNR). Highest identification had worried fear in pink background noise with 10.41%; most of the time, it was much lower or even at 0%.

A 2-dimensional *Non-metric Multi-Dimensional Scaling* (NMDS, [35, 36]) solution, computed in Matlab [33], is shown in Figure 4 for the clean and the 'worst' noisy, i. e., the pink -1 dB SNR (cf. Table 1), conditions. The goal of NMDS is a visual representation of the patterns of proximities. Starting with a random configuration of points, the pairwise distances between all points are calculated. The task is to find an optimal monotonic transformation of proximities (i. e., of the distances), in order to obtain optimally scaled data (disparities); the stress-value between the optimally scaled data (in a reduced dimensionality) and the distances has to be optimized by finding a new configuration of points. This step is iterated until a criterion is met. Normally, a 2-dimensional solution is interpreted as for meaningful dimensions and/or constellations.

As can be seen in Figure 4, the obtained stress and R^2 values are very strong, thus allowing for an interpretation of the resulting plots. When we mirror the pink plot at both x- and y-axis, we see that the 'cognitive-emotional space' represented is similar to the one for the clean condition. However, this space is more condensed – we speculate this is due to the higher confusion between classes in noisy conditions (cf. Figure 5). We can, roughly, identify the arousal dimension when drawing a line between SA and HO, and a (sort of) valence dimension when drawing a line between CO and EL. Valence is difficult to convey with acoustic means only, especially in nonsense utterances [37]; this might cause PA and PL being positioned side by side in spite of them having different valence (cf. Figure 2).

Previous studies indicate that the perception of linguistic content particularly deteriorates in white but not in pink noise [5]. Yet, our analysis shows that pink noise appears to have

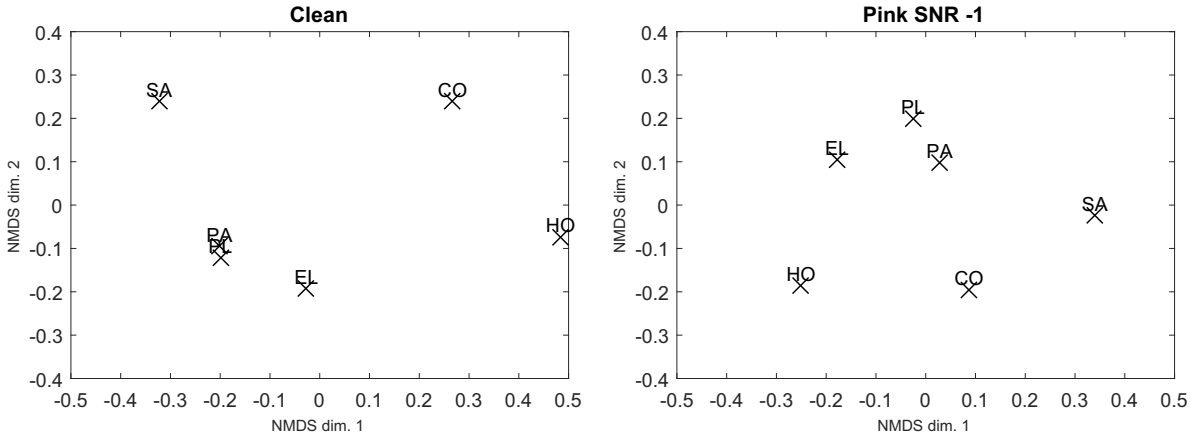

Figure 4: *2-dim. NMDS solution for clean (stress: 9.60E-07, R^2: 0.92) and pink -1 dB SNR (stress: 1.23E-16, R^2: 0.99) conditions.*

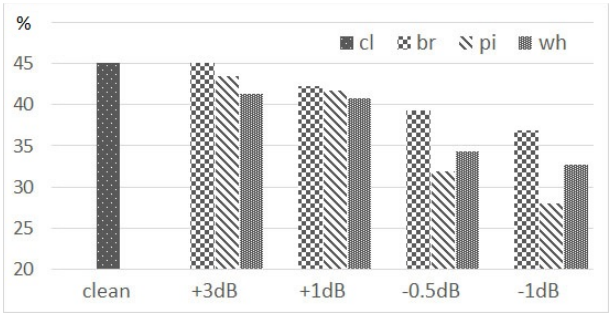

Figure 5: *Percentage of accuracy in the identification of emotions for each level and type of noise, and for the clean condition (36 utterances encoded in each bar).*

Table 3: *Percentage of hits in the recognition of male and female speakers in the clean and the three background noise conditions. Given is mean percentage of all emotions at -1 dB SNR (18 cases represented in each cell).*

%	cl	br	pi	wh
male	43.9	36.6	26.8	31.5
female	46.3	36.9	29.1	33.8

influenced emotion perception to a higher degree than white noise, brownian being in between white noise and clean condition (cf. Table 1). A reason might be that for 'normal' undistorted speech, the higher frequencies that are more masked in white noise than in the two other noises (cf. Figure 1) are more important, for instance, for recognising fricatives. In contrast, the low frequencies, most relevant for pitch and energy modulation, are more masked in brownian and pink noise. Pink noise generally masks more than brownian noise and displays the lowest performance.

When considering speaker gender, our analysis indicates that emotion perception is not related to the SNR variable: the mean percentage of accuracy (cf. Table 3) is lower in -1 dB SNR background noise conditions regardless of speaker gender. Furthermore, although brownian noise is characterized by a strong presence of low frequencies, no influence was shown on the perception of male speakers, being actually the less damaging of the three noises in the perception of emotional utterances produced by both genders.

The different intensities of noise influence perception of emotions in an ever increasing way for all the noises considered (cf. Figure 5). There seems to be no relationship between listeners' nationality and perception of emotional speech (due to the low number of different nationalities, this is no proof but a strong indication). Both groups of listeners (German and non-German) show similar percentages of accuracy for identifying emotion in speech, in different background noise conditions (cf. Table 4).

Table 4: *Percentage of accuracy in the recognition of emotional speech by German and non-German listeners in clean and background noise in -1 dB and +1 dB SNR.*

%	German	Non-German
cl	44.5	45.6
br $-1/+1\,(dB)$	36.3 / 42.1	37.3 / 42.3
pi $-1/+1\,(dB)$	28.1 / 40.7	27.9 / 42.6
wh $-1/+1\,(dB)$	33.4 / 39.1	31.9 / 42.4

5. Conclusion and outlook

Results presented in this paper indicate that noise reduces the performance of listeners in identifying emotions, pink noise being 'worst', and brownian noise being 'best'; this can be explained, in part, by looking at the spectral distribution of these noises. Interestingly, the 'cognitive-emotional space', especially the arousal dimension, is still maintained in noise. However, cold anger and especially sadness as low aroused emotions attract confusions to a considerable extent in noisier signals. This can be traced back to the masking of those acoustic (especially prosodic) characteristics in the other, higher aroused emotions, that are masked by the noises presented. Neither speaker gender nor culture (at least in our non-representative sample) seem to play a role.

Our future goal is to evaluate to which extent background noise from real environments can influence the perception of emotions. This will contribute to research areas related to human wellness and sonic environments [38], as for instance developing adaptive smart environment focus on the improvement of work conditions.

6. Acknowledgements

This work was supported by the European Union's Seventh Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu) and No. 688835 (RIA DE-ENIGMA). We would also like to thank all subjects for their participation.

7. References

[1] B. O. Olusanya, K. J. Neumann, and J. E. Saunders, "The global burden of disabling hearing impairment: A call to action," *Bulletin of the World Health Organization*, vol. 92, pp. 367–373, 2014.

[2] G. H. Bachara, J. Raphael, and W. J. Phelan, "Empathy development in deaf preadolescents," *American Annals of the Deaf*, vol. 125, pp. 38–41, 1980.

[3] K. M. Prashanth and V. Sridhar, "The relationship between noise frequency components and physical, physiological and psychological effects of industrial workers," *Noise and Health*, vol. 10, pp. 90–98, 2008.

[4] Y. Takata and A. K. Nábělek, "English consonant recognition in noise and in reverberation by Japanese and American listeners," *Journal of the Acoustical Society of America*, vol. 88, pp. 663–666, 1990.

[5] T. Shimizu, K. Makishima, M. Yoshida, and H. Yamagishi, "Effect of background noise on perception of english speech for japanese listeners," *Auris Nasus Larynx*, vol. 29, pp. 121–125, 2002.

[6] C. L. Rogers, J. J. Lister, D. M. Febo, J. M. Besing, and H. B. Abrams, "Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing," *Applied Psycholinguistics*, vol. 27, pp. 465–485, 2006.

[7] K. J. Van Engen and A. R. Bradlow, "Sentence recognition in native-and foreign-language multi-talker background noise," *Journal of the Acoustical Society of America*, vol. 121, pp. 519–526, 2007.

[8] M. Sereda, J. Davies, and D. A. Hall, "Pre-market version of a commercially available hearing instrument with a tinnitus sound generator: Feasibility of evaluation in a clinical trial," *International Journal of Audiology*, pp. 1–9, 2016.

[9] K. R. Scherer, S. Feldstein, R. N. Bond, and R. Rosenthal, "Vocal cues to deception: A comparative channel approach," *Journal of psycholinguistic Research*, vol. 14, pp. 409–425, 1985.

[10] M. Friend and M. J. Farrar, "A comparison of content-masking procedures for obtaining judgments of discrete affective states," *Journal of the Acoustical Society of America*, vol. 96, pp. 1283–1290, 1994.

[11] T. Johnstone and K. R. Scherer, "Vocal communication of emotion," in *Handbook of emotion*, M. Lewis and J. M. Haviland-Jones, Eds. New York: Guilford, 2000, vol. 2, pp. 220–235.

[12] K. R. Scherer, J. Koivumaki, and R. Rosenthal, "Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech," *Journal of Psycholinguistic Research*, vol. 3, pp. 269–285, 1972.

[13] K. R. Scherer, D. R. Ladd, and K. E. Silverman, "Vocal cues to speaker affect: Testing two models," *Journal of Language and Social Psychology*, vol. 5, pp. 1346–1356, 1984.

[14] B. Schuller, D. Arsić, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," in *Proc. Int. Conf. on Speech Prosody*, Dresden, Germany, 2006, pp. 276–289.

[15] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 4, Honolulu, HY, 2007, pp. 941–944.

[16] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, pp. 1062–1087, 2011.

[17] T. Bänziger, H. Pirker, and K. Scherer, "GEMEP-GEneva Multimodal Emotion Portrayals: A corpus for the study of multimodal emotional expressions," in *Proc. Conf. on Language Resources and Evaluation*, Genova, Italy, 2006, pp. 15–19.

[18] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, pp. 5–32, 2003.

[19] P. Ekman, "Expression and the nature of emotion," *Approaches to Emotion*, vol. 3, pp. 19–344, 1984.

[20] R. Plutchik, "Emotions in early development: A psychoevolutionary approach," *Emotion: Theory, Research, and Experience*, vol. 2, pp. 221–257, 1983.

[21] ——, *The emotions: Facts, theories, and a new model*. Lanham, Maryland: University Press of America, 1991.

[22] A. Ortony and T. J. Turner, "What's basic about basic emotions?" *Psychological Review*, vol. 97, pp. 315–331, 1990.

[23] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[24] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, pp. 227–256, 2003.

[25] C. Oflazoglu and S. Yildirim, "Recognizing emotion from turkish speech using acoustic features," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 26, pp. 1–11, 2013.

[26] I. R. Murray and J. L. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Communication*, vol. 16, pp. 369–390, 1995.

[27] K. R. Scherer, R. Banse, and H. G. Wallbott, "Emotion inferences from vocal expression correlate across languages and cultures," *Journal of Cross-Cultural Psychology*, vol. 32, pp. 76–92, 2001.

[28] H. A. Elfenbein and N. Ambady, "On the universality and cultural specificity of emotion recognition: A meta-analysis," *Psychological Bulletin*, vol. 128, pp. 203–235, 2002.

[29] D. Sauter, "An investigation into vocal expressions of emotions: The roles of valence, culture, and acoustic factors," Ph.D. dissertation, University College London, 2006.

[30] D. A. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations," in *Proc. of the National Academy of Sciences*, vol. 107, 2010, pp. 2408–2412.

[31] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. Int. Conf. and Workshops on Automatic Face and Gesture Recognition*, Shanghai, China, 2013, pp. 1–8.

[32] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proc. Interspeech*, Lyon, France, 2013, pp. 148–152.

[33] I. Mathworks, "Matlab: R2014a," *Mathworks Inc, Natick*, 2014.

[34] S. Hantke, F. Eyben, T. Appel, and B. Schuller, "iHEARu-PLAY: Introducing a game for crowdsourced data collection for affective computing," in *Proc. Int. Workshop on Automatic Sentiment Analysis in the Wild held in conjunction with the biannual Conference on Affective Computing and Intelligent Interaction*, Xi'an, China, 2015, pp. 891–897.

[35] J. B. Kruskal, "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, vol. 29, pp. 115–129, 1964.

[36] J. Kruskal and M. Wish, *Multidimensional Scaling*. Beverly Hills and London: Sage University, 1978.

[37] J.-A. Bachorowski and M. J. Owren, "Vocal expressions of emotion," in *Handbook of emotions*, M. Lewis, J. M. Haviland-Jones, and L. F. Barret, Eds. New York, USA: The Guilford Press, 2008, pp. 196–210.

[38] E. Parada-Cabaleiro, A. Baird, N. Cummins, and B. Schuller, "Stimulation of psychological listener experiences by semi-automatically composed electroacoustic environments," in *Proc. Int. Conf. on Multimedia and Expo (ICME 2017)*, Hong Kong, China, 6 pages.