Speech Prosody 2018
13-16 June 2018, Poznań, Poland

The Effect of Noise on Emotion Perception in an Unknown Language

Odette Scharenborg[1], *Sofoklis Kakouros*[2], *and Jiska Koemans*[1]

[1] Centre for Language Studies, Radboud University Nijmegen, The Netherlands
[2] Department of Signal Processing and Acoustics, Aalto University, Finland
o.scharenborg@let.ru.nl, sofoklis.kakouros@aalto.fi, jiska.koemans@student.ru.nl

Abstract

This is the first study investigating the influence of realistic noise on verbal emotion perception in an unknown language. We do so by linking emotion perception to acoustic characteristics known to be correlated with emotion perception and investigating the effect of noise on the perception of these acoustic characteristics. Dutch students listened to Italian sentences in five emotions and were asked to indicate the emotion that was conveyed in the sentence. Sentences were presented in a clean and two babble noise conditions. Results showed that the participants were able to recognise emotions in the, for them, unknown language, and continued to perform above chance even in fairly bad listening conditions, indicating that verbal emotion may contain universal characteristics. Noise had a similar detrimental effect on the perception of the different emotions, though the impact on the use of the acoustic parameters for different emotion categories was different.

Index Terms: verbal emotion perception, unknown language, background noise, acoustic parameters, universals

1. Introduction

In today's multilingual society, many people regularly communicate in a language other than their native language, and typically in an environment with background noise. This combination of imperfect language knowledge and a degraded signal is known to hamper sound processing (for a review: [1]), word recognition (e.g., [2]), and also the uptake and use of prosodic features [3], although prosodic information seems to survive the masking effect of background noise to a certain extent [4]. Despite emotion perception being an important aspect of human communication and social interaction [5], little is known about the impact of background noise on emotion perception (but see [6][7] for emotion perception with simulated hearing loss). To our knowledge, only one study investigated the effect of noise on (human) emotion perception [8] (although it is a known problem in the field of ASR, e.g., [9]), showing the negative impact of white, pink, and brownian noise on emotion perception.

We investigate the influence of babble noise on verbal emotion perception in an unknown language, thus extending the research on emotion perception in noise to realistic noise. Where [8] tested listeners with different language backgrounds on nonsense sentences, we test listeners with the same native language (Dutch) on an existing (for them) "unknown" language, i.e., Italian, thus controlling for possible language pair influences. Importantly, because different emotions are conveyed using different acoustic parameters [10], we link emotion perception to acoustic characteristics known to be correlated with emotion perception and thus, for the first time, investigate the effect of noise also on the perception of these acoustic characteristics for emotion perception.

We chose seven acoustic parameters that have most often been found to correlate with the five emotions used in the current study (e.g., [6][7] [11][12]): Intensity range, Mean intensity, mean F0, F0 range, F0 variance, the Hammarberg index, and the slope of the long-term average spectrum (slope LTAS; [13]). In addition, we also used the spectral slope derived from the computation of MFCCs [14][15]. These factors were included as factors to the statistical analysis to investigate which acoustic characteristics are used for emotion perception and survive the noise.

2. Experimental set-up

2.1. Participants

Twenty-four native Dutch listeners (4 males; mean age=23.0, SD=4.2), recruited from the Radboud University subject pool, participated in the experiment. None of the participants reported a history of language, speech, or hearing problems. The participants were paid for their participation.

2.2. Materials

The stimuli were taken from the Italian EMOVO corpus [16],[17], which consists of 588 acted Italian emotional utterances portraying six emotions and a neutral state. These emotions are *anger*, *fear*, *sadness*, *joy*, *surprise*, and *disgust*. Six professional actors (age 23-30 years; 3 males and 3 females) enacted the emotions in 14 emotionally neutral sentences: nine were semantically neutral (e.g., 'workers get up early') and five were 'nonsense'[1] sentences with correct grammar. For the current experiment, a subset of in total 100 recordings was selected, consisting of ten different (5 semantically neutral and 5 nonsense) sentences, each portrayed in four emotions and a neutral state (henceforth collectively referred to as *emotions* for simplicity) and each produced by a male and a female speaker. Therefore, all sentences occurred in all emotions. Similar to [31][32][33][10], the emotions chosen for this experiment were *anger*, *fear*, *sadness*, *joy*, and *neutral*. As the corpus is based on the categorical model of emotions, we used the categorical approach in our emotion perception experiment.

Previous experiments on emotion perception in EMOVO showed variety in how well the different actor's intended emotions were perceived by Italian listeners [17][18][19]. For the current study, the female and male actor who were found to portray a specific emotion best according to [18][19] were

[1] Statistical analyses showed no significant differences between the normal and the nonsense utterances (nor were these expected as our listeners do not know Italian). This factor was therefore not included in the statistical analyses reported here.

10.21437/SpeechProsody.2018-74

chosen to represent that emotion in the subset used in this study. For *anger* and *joy* these were speakers f2 and m2, for *sadness* speakers f1 and m1, for *fear* f2 and m1, for *neutral* f3 and m3.

The sentences were presented in a clean and two noise conditions. For the noise conditions, after normalisation of intensity, Italian 8-speaker babble noise was added to the sentences at two different SNRs, i.e., SNR -5 dB and SNR +2 dB using a custom-made *Praat* [20] script. To create the 8-speaker babble, 8 Italian sentences, spoken by 8 different speakers (4 males and 4 females) from a subset of a different Italian corpus [21] were randomly selected. After normalising the intensity level, the 8 sentences were mixed together. Each sentence was preceded by 200 ms of leading noise and followed by 200 ms of trailing noise. A Hamming window was applied to the noise, with a fade in / fade out of 10 ms.

The SNRs were determined on the basis of an earlier study [22], and chosen such that for the easier SNR, the noise was distracting but the participants could still relatively easily hear the sentence, while the lowest SNR was chosen such that hearing the sentence became quite difficult but not impossible.

2.3. Experimental lists

Each participant was randomly assigned to one of 12 different experimental lists. Each list consisted of 120 sentences, with 40 sentences for each of the clean/noise listening conditions. Each subset (i.e., listening condition) contained eight utterances for each of the five emotions: four utterances each from the selected female and male speakers. Per speaker, two utterances were of semantically normal content and two were nonsense utterances. Each sentence-emotion-speaker stimulus appeared only once per 40-item subset, and each sentence-emotion-speaker-listening-condition stimulus appeared only once per experimental list. Sentences within each subset were randomised. The order of the listening conditions was counterbalanced across listeners.

2.4. Procedure

Participants were tested individually in a sound-treated booth. The stimuli were presented over closed headphones at a comfortable sound level. Participants were asked to determine for each of the 120 utterances which of the five emotions they thought it portrayed. They were instructed not to pay attention to the content of the utterances but to focus on the emotion with which the sentence was uttered. The five emotions (*anger*, *fear*, *sadness*, *joy*, and *neutral*) and an "I don't know option" were presented on a computer screen and for each of the emotions a corresponding key on the keyboard was labelled. The participants had to press the key representing the emotion they thought they heard. Every utterance was played once.

2.5. Acoustic parameters

The utterances were initially downsampled from 44.1kHz to 16kHz and seven acoustic features were extracted that are known to correlate with the five emotions in the current study (e.g., [6][7][10][12]): Intensity range, Mean intensity, mean F0, F0 range, F0 variance, the Hammarberg index, and the slope of the long-term average spectrum (LTAS). Since there is no standard way to quantify the relative contribution of high versus low frequency bands of the spectrum and different measures may carry complementary information that can be relevant for emotion perception an additional measure was used for spectral slope that was derived from MFCC computation (C1 - see, e.g., [15], for more details). All acoustic features were computed using a 25ms window with 10ms hop size. Specifically, intensity was computed as the temporal energy of the speech signal (see, e.g., [23]), F0 using the YAAPT algorithm [24], the Hammarberg index by taking the difference in dB between the maximum spectral energy in the 0-2kHz band and maximum energy in the 2-5kHz band [12][25], and the slope of the LTAS by fitting a first order polynomial into the averaged magnitude spectrum. Finally, the MFCC-based slope measure was computed by taking the first MFCC (C1) for each frame [15].

Following the computation of the raw feature values, feature aggregates were computed at the utterance level. For F0, the mean, range (maximum-minimum of the feature during the utterance), and variance of the F0 estimates in Hz were computed (note that normalisation was not applied in this case – see, e.g., [7]). Intensity measures were computed from the logarithmically normalised (in dB) signal energy. Specifically, the range and mean of intensity for each utterance were computed. For the Hammarberg index and the MFCC derived spectral slope, the mean per utterance feature value was measured and, finally, LTAS was computed across each utterance and the final slope measure was taken as the direction of the regression line (see, e.g., [13] for a similar approach).

3. Results

Figure 1 shows the proportion of utterances per emotion that were correctly identified for each of the three noise conditions. Statistical analyses on the accuracy of the recognised emotions were carried out using generalised linear mixed-effect models (e.g., [26]), containing fixed and random effects. To obtain the final, best-fitting model containing only statistically significant effects, we used the backward stepwise selection procedure as, e.g., described in [27]. The dependent variable was correct ('1') or incorrect ('0') emotion recognition. Fixed factors were Emotion (*anger* on the intercept; nominal variable), Noise (clean on the intercept; nominal variable), and Gender. Sentence, Speaker (only in the overall analysis), and Subject were entered as random factors. Random by-sentence, by-speaker, and by-subject slopes for Noise were added and tested through model comparisons in all analyses.

Table 2 displays the estimates of the fixed effects in the best-fitting model (due to space limitations the non-significant interactions between Emotion and Noise are not shown, all p>.26). Significantly fewer correct answers (see also Figure 1) were given for *fear*, *sadness*, and *joy* compared to *anger*. Moreover, important to our research question, when noise was present, for significantly fewer utterances the emotions were correctly identified compared to the clean listening condition. Note that *neutral* only showed a significant deterioration in accuracy for SNR -5 dB compared to the clean condition. Although there are large differences in the accuracy in perceiving the five emotions, in all cases, the Dutch listeners performed (well) above chance level in all listening conditions.

Subsequently, the acoustic parameters were added to the analysis as main effects and in interaction with Emotion, Gender, and Noise. The results of this overall analysis showed significant interactions between several acoustic parameters and several of the emotions. Intensity range was the only acoustic feature to predict overall emotion perception: significantly fewer correct answers were given for increasing Intensity range (β=-.633, SE=.177, p=<.001). Moreover, significant interactions between different emotions and the Hammarberg Index (*joy* and *neutral*; ps < .039), Slope MFCC (*fear*, ps < .026), Slope LTAS (*sadness*, p = .023), F0 variance (*joy*, p = .001), and mean F0 (*sadness*, p = .043) were found, as well as significant interactions between the two SNR levels and Slope MFCC (SNR -5, p = .009), Intensity range (SNR 2, p = .009), and mean F0 (SNR 2, p = .044; SNR -5, p=.073).

Table 1. *Fixed effect estimates for fixed effects in the best-fitting models of performance for the acoustic factor analysis per emotion.*

Acoustic factor	*fear*			*joy*			*anger*			*neutral*			*sadness*		
	β	*SE*	*p*	β	*SE*	*p*	β	*SE*	*p*	β	*SE*	*p*	β	*SE*	*p*
Intercept	-1.680	1.286	.187	-.472	.543	.384	.735	1.654	.657	1.796	.496	**<.001**	.441	1.182	.710
Noise: +2 dB	-.710	.312	**.023**	-.954	.252	**<.001**	-1.108	.356	**.002**	1.522	.571	**.008**	.178	1.262	.888
Noise: -5 dB	-1.486	.321	**<.001**	-2.250	.294	**<.001**	-1.666	.349	**<.001**	-.688	.398	.084	-4.079	1.272	**.001**
Gender	4.924	1.775	**.006**	5.012	1.133	**<.001**	-13.303	3.789	**<.001**	16.061	10.303	.119	7.939	4.964	.110
Mean F0	2.447	1.141	**.032**	1.708	.420	**<.001**	1.693	.926	.067	-.881	1.209	.466	-2.825	1.585	.075
F0 range	-1.566	.653	**.016**	1.195	.456	**.009**	-.291	1.142	.799				1.524	.773	**.049**
F0 variance	2.483	.999	**.013**	1.354	.435	**.002**	-.338	.961	.725	.872	.732	.233	-2.890	1.653	.080
Mean intensity							-2.290	.992	**.021**	1.581	.599	**.008**	.045	.488	.927
Intensity range				-.309	.323	.340	-.644	.239	**.007**				.957	1.934	.621
Hammarberg index				-.586	.175	**<.001**	1.650	.644	**.010**	-.460	.281	.102	.332	.432	.441
Slope MFCC	.568	.243	**.020**				-3.400	1.076	**.002**				-.877	.516	.089
Slope LTAS							-1.560	.918	.089						

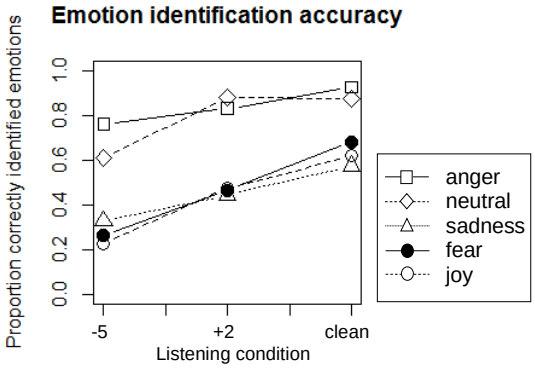

Figure 1. *Proportion of correctly identified utterances for each of the five emotions and three noise conditions separately.*

Table 2. *Fixed effect estimates for the best-fitting model of the overall accuracy analysis, n=3000.*

Fixed effect	β	*SE*	*p*
Intercept	2.194	.293	< .001
Emotion: *fear*	-1.185	.335	< .001
Emotion: *sadness*	-1.188	.329	< .001
Emotion: *joy*	-1.460	.332	< .001
Emotion: *neutral*	-.311	.368	.397
SNR +2 dB	-.962	.334	.004
SNR -5 dB	-1.420	.322	< .001
Gender: male	.946	.242	< .001
Emotion: *neutral* × SNR: +2 dB	1.009	.454	.026
Emotion: *fear* × Gender: male	-1.386	.300	< .001
Emotion: *sadness* × Gender: male	-.946	.296	.001
Emotion: *joy* × Gender: male	-1.387	.301	< .001
Emotion: *neutral* × Gender: male	-.696	.323	.031

To investigate the use of the various acoustic parameters on emotion perception and the influence of noise thereof in more detail, separate analyses were carried out for the five emotions. Table 1 displays the estimates of the fixed effects Noise, Gender, and the acoustic parameters under investigation (those significant are indicated in bold for ease of reading) for the separate emotion analyses. Due to space limitations, the significant interactions are described below rather than added to Table 1. As expected, for all emotions, the presence of noise had a significantly detrimental effect compared to the clean condition, except for the SNR +2 condition for *sadness* and the SNR -5 condition for *neutral*. Table 1 only lists the factors which are significant or are part of significant interactions with SNR and/or Gender. What is immediately clear is that for all emotions, many different acoustic features are associated with emotion perception, and all acoustic factors are associated with the perception of multiple emotions.

Fear: A higher slope MFCC and a smaller F0 range were associated with more correct *fear* recognitions. Moreover, a higher Mean F0 was associated with more correct *fear* recognitions and even more so for the male speaker (β=5.264, *SE*=1.457, *p*<.001). A larger F0 variance was associated with more correct *fear* recognitions for the female speaker while the opposite was true for the male speaker (β=3.959 *SE*=1.490, *p*=.008). The use of these acoustic parameters was however not impeded by the presence of noise as shown by the lack of an interaction between these parameters and Noise.

Joy: A lower Hammarberg Index and higher Mean F0, F0 range, and F0 variance were associated with more correct *joy* responses. An interaction between Noise: -5 dB and Mean F0 (β=-1.129, *SE*=.373, *p*=.003) shows that in the two easiest listening conditions, a higher Mean F0 was associated with an increase in correct *joy* responses, while listeners were no longer able to use Mean F0 in the most difficult listening condition. Significant interactions between both Noise conditions and F0 range (+2dB: β=.954, *SE*=.329, *p*=.004; -5dB: β=1.297, *SE*=.383, *p*<.001) indicate that when noise was present, a higher F0 range was associated with an increase in correct *joy* responses. Listeners thus relied more on F0 range when listening conditions deteriorated for the correct recognition of *joy*. Finally, a higher Intensity range predicted more correct *joy* responses for the male speaker (β=4.340, *SE*=1.133, *p*<.001).

Anger: A higher Mean intensity and Intensity range were associated with significantly fewer correct responses, but this was less the case for the male speaker (Mean intensity: β=3.406, *SE*=1.583, *p*=.031; Intensity range: β=36.597, *SE*=9.368, *p*<.001). A higher Slope MFCC was associated with significantly fewer correct responses. The use of these acoustic parameters was however not impeded by the presence of noise. A higher Hammarberg Index was associated with significantly more correct responses, but this was significantly less so for the male speaker (β=-7.607, *SE*=1.923, *p*<.001). For the male speaker (but not the female speaker) F0 range (β=-2.859, *SE*=1.456, *p*=.050) and Mean F0 (β=-7.916, *SE*=2.463, *p*=.001) were associated with relatively more correct *anger* responses, while the reverse was the case for F0 variance (β=2.831, *SE*=1.233, *p*=.022). Finally, again for the male speaker only, a higher Slope LTAS was associated with relatively fewer correct responses (β=-9.976, *SE*=2.207, *p*<.001).

Neutral: In the clean, a higher Mean intensity is associated with more correct *neutral* responses, while this was less the case in the presence of noise (+2dB: β=-2.939, *SE*=.887 *p*<.001; -5

dB: β=-2.133, *SE*=.682, *p*=.002), indicating that the listeners were no longer able to correctly use Mean intensity for the recognition of *neutral*. Moreover, for the male speaker (but not the female speaker), a higher Hammarberg Index (β=2.246, *SE*=.868 *p*=.010) and Mean F0 (β=27.695, *SE*=12.996, *p*=.033) were associated with relatively more correct responses and a higher F0 variance with relatively fewer correct responses (β= -20.491, *SE*=8.011, *p*=.011).

Sadness: A higher F0 range was associated with more correct responses but less so for the male speaker (β=-4.719, *SE*=2.2374, *p*=.047). Moreover, while Mean F0 was not used as a cue for *sadness* detection, in the presence of noise a higher Mean F0 was associated with fewer correct responses (+2dB: β=-1.520, *SE*=.456; *p*<.001; -5dB: β=-1.123, *SE*=.465; *p*=.016), indicating that listeners started to use Mean F0 for *sadness* detection when listening conditions deteriorated. Finally, for the male speaker (but not the female speaker), the Hammarberg Index (β=-3.000, *SE*=.874; *p*<.001) and Mean intensity (β=-1.442, *SE*=.657; *p*=.028) were associated with relatively fewer correct responses while the opposite pattern was found for Slope MFCC (β=3.895, *SE*=1.087 *p*<.001) and F0 variance (β=25.068, *SE*=9.672; *p*=.010).

4. Discussion

The current study showed that Dutch listeners are able to perceive emotions in a language that is unknown to them, i.e., Italian. Although there are large differences in accuracy in perceiving the five different emotions, the Dutch listeners performed (well) above chance level. *Anger* and *neutral* were easiest to recognise (with a low confusability of *anger* with the other four emotions). This is in line with [32],[33] who also found that *anger* and *neutral* were among the best recognised emotions, using different languages and different listener groups. *Sadness*, *joy*, and *fear* were (much) less well recognised by our listener group. The difficulty of recognising *joy* and *fear* is again in line with [32],[33]. The biggest difference between our results and those of others concerns *sadness*, which in other studies is often among the best-recognised emotions [8][31][32][33].

The results thus showed that listeners without any knowledge of a language can perceive verbal emotions in that language. *Nonverbal* emotion is said to contain specific universal characteristics, i.e., some emotions are fairly consistent across languages and societies [28], making it possible to recognise emotions across different cultures (e.g., [29]). Our findings add to existing findings (e.g., [30][31][32][33]) showing that *verbal* emotion may also contain universal characteristics. Nevertheless, which emotions are easiest to recognise seems to depend on several factors, including the language (pair) and listener group. Nevertheless, listeners are better in correctly perceiving emotions when they are listening in their native language, indicating that there are also language-dependent characteristics of verbal emotion perception [31][32][33].

Background noise negatively influenced verbal emotion perception in an unknown language, with basically no differences of the effect of noise on the different emotion categories, although *anger* perception was less influenced by mild noise. These results are in line with those reported for native listeners by [8] who also found little difference between emotions, with the exception that in their case the perception of *sadness* was less influenced by the presence of noise.

Comparing the acoustic parameters used for the recognition of the different emotion categories in Table 1 shows that for the perception of some emotions many different acoustic cues are used (*fear*, *joy*, and *anger*) while for others only one acoustic cue is used (*neutral* and *sadness*). Interestingly, the number of acoustic cues that is used for the perception of an emotion does not seem to correlate with the effect of background noise on the perception of that emotion.

We observed a differential effect of noise on the use of the different acoustic parameters for emotion perception. Noise did not interfere with the use of the acoustic parameters conveying *fear* and *anger*. However, the presence of noise meant that Mean intensity was no longer correctly used for the recognition of *neutral*, and Mean F0 could no longer be used for the recognition of *joy* when noise was present at an SNR of -5 dB. Reversely, when noise was present, F0 range was used more for *joy* recognition, and Mean F0 was used more for the recognition of *sadness*. Surprisingly, the effect of noise on the use of the *same* acoustic parameter was dependent on the emotion to be recognised. A more detailed analysis of the acoustic parameters for each of the emotion categories and for each of the actors and a comparison with those of the background noise will be carried out to further investigate this differential effect of noise on the acoustic parameters.

The stimuli in this study consisted of acted emotional speech, which may be more extreme and prototypical than natural speech [34],[35]. Prototypical, acoustic patterns for emotions might be easier to perceive than natural emotional speech, which means that the here presented results may be an upper-bound for emotion perception in an unknown language.

In the emerging field of affective computing, correct emotion perception is important for the improvement of man-machine-interaction-systems and multi-media processing [36]. Studies in this area have shown that also for automatic systems, the presence of noise interferes with verbal emotion perception (e.g., [37]). For speech recognition, humans have been found to outperform machines in the presence of noise by a large margin [38]. It is however difficult to compare the degree of impact of noise on the human listeners in the current study with those on machines as different types of noise (babble vs. white noise in [37]) have been used in the different studies. More systematic studies are needed to clarify this issue.

This is the first paper, to our knowledge, that uses realistic background noise to investigate the effect of background noise on verbal emotion perception. The results showed that listeners without any knowledge of a non-native language can perceive verbal emotions in that language, and they continue to perform above chance level even in fairly bad noise conditions. The presence of noise was found to have an increasingly detrimental effect on verbal emotion perception, and impacted the use of different acoustic characteristics for emotion perception differently depending on the emotion category. No clear correlation between the effect of noise on acoustic cue use and the perception of the specific emotion was found. Future research will have to investigate this finding more closely.

5. Acknowledgements

This work was carried out by J. K. as part of an intership under the supervision of O. S., and was sponsored by a Vidi-grant from NWO (grant number: 276-89-003) to O. S. The authors thank Martin Cooke for creating the babble noise files, and Marthe Scholten for help running the experiment.

6. References

[1] M.L.G. Garcia Lecumberri, M. Cooke, and A. Cutler, "Non-native speech perception in adverse conditions: A review", *Speech Communication*, vol. 52, pp. 864-886, 2010.

[2] O. Scharenborg, J. Coumans, R. van Hout, "The effect of background noise on the word activation process in non-native spoken-word recognition", Journal of Experimental Psychology: Learning, Memory, and Cognition, 2017. doi:10.1037/xlm0000441.

[3] O. Scharenborg, E. Kolkman, S. Kakouros, and B. Post, "The effect of sentence accent on non-native speech perception in noise", *Interspeech*, pp. 863-867, 2016.

[4] M. Van Zyl and J.J. Hanekom, "Speech perception in noise: A comparison between sentence and prosody recognition", *Journal of Hearing Science, 1*(2), pp. 54-56, 2011.

[5] R.J.R. Blair, "Facial expressions, their communicatory functions and neuro- cognitive substrates", *Philos. Trans. R. Soc. Lond. Biol. Sci*. 358, 561–572, 2003.

[6] M. Chatterjee, D. Zion, M.L. Deroche, B. Burianek, C. Limb, A. Goren, A.M. Kulkarni, and J.A. Christensen, "Voice emotion recognition by cochlear-implanted children and their normally-hearing peers", *Hear Res.*, vol. 322, pp. 151-162, 2015.

[7] X. Luo, Q. Fu, and J.J. Galvin, "Cochlear implants special issue article: vocal emotion recognition by normal-hearing listeners and cochlear implant users", *Trends Amplif.*, vol. 11, no. 4, pp. 301-315, 2007.

[8] E. Parada-Cabaleiro, A. Baird, A. Batliner, N. Cummins, S. Hantke, B.W. Schuller, "The perception of emotions in noisified nonsense speech", Proceedings of Interspeech, 2017.

[9] 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.

[10] C. Sobin and M. Alpert, "Emotion in speech: The acoustic attributes of tear, anger, sadness and joy", *Journal of Psycholinguistic Research*, vol. 28, no. 4, pp. 347-365, 1999.

[11] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression", *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, Mar. 1996.

[12] J. Schmidt, E. Janse, E., and O. Scharenborg, "Perception of emotion in conversational speech by younger and older listener", *Frontiers in Psychology, section Language Sciences*, vol. 7, pp. 781, 2016.

[13] M. Guzman, S. Correa, D. Muñoz, and R. Mayerhoff, R., "Influence on spectral energy distribution of emotional expression", *Journal of Voice*, vol. 27, no. 1, pp. 129, 2013.

[14] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, "Spectral moment features augmented by low order cepstral coefficients for robust ASR", *IEEE Signal Processing Letters*, *17*(6), pp. 551-554, 2010.

[15] S. Kakouros, O. Räsänen, and A. Paavo, "Evaluation of Spectral Tilt Measures for Sentence Prominence Under Different Noise Conditions," *Interspeech*, Stockholm, Sweden, 3211-3215, 2017.

[16] Corpus downloaded from: http://voice.fub.it/activities/corpora/emovo/index.html

[17] G. Constantini, I. Iadarola, A. Paoloni, and M. Todisco, "EMOVO corpus: an Italian emotional speech database", *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 2014.

[18] C. Giovannella, D. Conflitti, and R. Santoboni, "Transmission of vocal emotion: Do we have to care about the listener? The case of the Italian speech corpus EMOVO", *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1-6, 2009.

[19] C. Giovannella, D. Floris, and A. Paoloni, "An exploration on possible correlations among perception and physical characteristics of EMOVO emotional portrayals", *Interaction Design and Architecture(s) Journal*, 10, pp. 102-111, 2012.

[20] P. Boersma and D. Weenink, D. "Praat: doing phonetics by computer [Computer program]", 2013. Retrieved from http://www.praat.org/

[21] Corpus downloaded from: http://www.clips.unina.it/en/index.jsp

[22] J. Koemans, "Emotion perception in adverse listening conditions", *BA thesis*¸ Department of Linguistics, Radboud University, Nijmegen, The Netherlands, 2016.

[23] S. Kakouros, and O. Räsänen, "3PRO-An unsupervised method for the automatic detection of sentence prominence in speech", *Speech Communication*, 82, 67-84, 2016.

[24] S.A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking", *The Journal of the Acoustical Society of America, 123*(6), 4559-4571, 2008.

[25] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities", *Acta Otolaryngol.*, vol. 90, 441-451, 1980.

[26] R. H. Baayen, D. J. Davidson, and D. M. Bates, "Mixed-effects modeling with crossed random effects for subjects and items", *Journal of Memory and Language*, vol. 59, pp. 390-412, 2008.

[27] O. Scharenborg, A. Weber, and E. Janse, "The role of attentional abilities in lexically-guided perceptual learning by older listeners", *Attention, Perception, and Psychophysics*, vol. 77, no. 2, pp. 493-507, 2015.

[28] D. Sauter, F. Eisner, P. Ekman, S. Scott, "Universal vocal signals of emotion", *Proceedings of Cognitive Science*, pp. 2251-2255, 2009.

[29] P. Ekman, "An argument for basic emotions", *Cognition and Emotion*, vol. 6, pp. 169-200, 1992.

[30] W. Da Silva, P.A. Barbosa, and A. Abelin, "Cross-cultural and cross-linguistic perception of authentic emotions through speech: An acoustic–phonetic study with Brazilian and Swedish listeners", *Delta*, vol. 32, no. 2, 2016.

[31] W.F. Thompson and L.-L. Balkwill, "Decoding speech prosody in five different languages", *Semiotica*, vol. 158, no. ¼, pp. 407-424, 2006.

[32] 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, no. 1, pp. 76-92, 2001.

[33] M.D. Pell, L. Monetta, and S. Paulmann, "Recognizing emotions in a foreign language", *Journal of Nonverbal Behaviour*, vol. 33, pp. 107-120, 2009.

[34] K.R. Scherer, "Vocal affect expression: a review and a model for future research", *Psychol. Bull.*, vol. 99, pp. 143-165, 1986.

[35] J. Wilting, E. Krahmer, and M. Swerts, "Real vs. acted emotional speech", in *Proceedings of the 9th International Conference on Spoken Language Processing*, pp. 805-808, 2006.

[36] E. Shriberg, "Spontaneous speech: How people really talk and why engineers should care", *Interspeech*, pp. 1781-1784, 2005.

[37] B. Schuller, D. Arsić, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets", *Speech Prosody*, 2006.

[38] R. Lippmann, "Speech recognition by machines and humans", *Speech Communication*, vol. 22, no. 1, 1-15, 1997.