Misperceptions of the emotional content of natural and vocoded speech in a car

Jaime Lorenzo-Trueba[1], *Cassia Valentini-Botinhao*[2], *Gustav Eje Henter*[1], *Junichi Yamagishi*[1,2]

[1]National Institute of Informatics, Tokyo, Japan
[2]The University of Edinburgh, Edinburgh, UK

jaime@nii.ac.jp,cvbotinh@inf.ed.ac.uk,gustav@nii.ac.jp,jyamagish@nii.ac.jp

Abstract

This paper analyzes a) how often listeners interpret the emotional content of an utterance incorrectly when listening to vocoded or natural speech in adverse conditions; b) which noise conditions cause the most misperceptions; and c) which group of listeners misinterpret emotions the most. The long-term goal is to construct new emotional speech synthesizers that adapt to the environment and to the listener. We performed a large-scale listening test where over 400 listeners between the ages of 21 and 72 assessed natural and vocoded acted emotional speech stimuli. The stimuli had been artificially degraded using a room impulse response recorded in a car and various in-car noise types recorded in a real car. Experimental results show that the recognition rates for emotions and perceived emotional strength degrade as signal-to-noise ratio decreases. Interestingly, misperceptions seem to be more pronounced for negative and low-arousal emotions such as calmness or anger, while positive emotions such as happiness appear to be more robust to noise. An ANOVA analysis of listener meta-data further revealed that gender and age also influenced results, with elderly male listeners most likely to incorrectly identify emotions.

Index Terms:emotional perception, speech in noise, emotion recognition, car noise

1. Introduction

Speech synthesis systems have progressed remarkably recently, and we are now becoming capable of generating natural-sounding synthetic speech almost comparable to human speech: convolutional neural networks-based approaches [1], waveform modeling approaches [2, 3] and generative-adversarial network based approaches [4] have proven to be extremely successful for generating very high-quality synthetic speech (and other types of sounds such as onomatopoeia or music). We have also proved that synthetic speech generated by neural network-based systems can be adaptable and controllable in terms of the perceived age, gender, speaker identity [5] and also capable of supporting multiple languages [6].

One aspect that is still under development and hence requires fundamental research is the generation of speech in adverse environments such as noisy or reverberant conditions. It is known that speech intelligibility degrades significantly under noisy conditions [7, 8], becoming this effect even more significant with age [9]. Human communication implicitly compensates this degradation making use of Lombard speech [10]. There are also several attempts to use environmental noise information to improve the intelligibility of speech synthesizers adaptively [11, 12].

In this paper we hypothesize that not only speech intelligibility, but also interpretation of the emotional content of an utterance may be recognized incorrectly when listening to natural or synthetic speech in adverse conditions. We also hypothesize that this degradation in emotional recognition capabilities and perceived emotional strength may depend on the listener's age and gender. To validate the hypothesis this paper analyzes: a) how often listeners interpret the emotional content of an utterance incorrectly when listening to natural or vocoded speech in adverse conditions; b) which noise conditions cause the most misperceptions; and c) which group of listeners misinterpret emotions the most. If the experimental results underpins our hypothesis, highly expressive emotional speech synthesizers will need to compensate not only intelligibility but also for the emotional content of synthetic speech to target specific groups of listeners under the varied adverse conditions. That is, there will be a need for emotional speech synthesizers that adapt to the environment and to the listener, which is our long-term scientific goal.

We performed a large-scale listening test where over 400 crowd-sourced listeners between the ages of 21 and 72 assessed natural and vocoded acted emotional speech stimuli. In emotional speech synthesis, the acted emotions are typically utilized for acoustic modelling. Hence in our experiment we have used 7 acted emotional speech uttered by a professional voice actress. Vocoded speech is used as a proxy for text-to-speech based systems that have slightly worse quality than natural speech. The stimuli were then artificially degraded using a room impulse response recorded in a car [13] and various in-car noise types recorded in the same car . The listeners were asked to identify emotional content of speech utterances under adverse conditions and to rate perceived strength of the emotional content. We have analyzed emotional recognition rates and perceived emotional strength with respect to SNRs, emotional categories, and listeners' information.

The paper is structured as follows: section 2 introduces the emotional speech corpus that we have used for the evaluation, then section 3 explains how we recorded the noise to be added to the emotional speech and the exact procedure in which it was added. In section 4 we describe the evaluation process, showing the results in section 5. Finally in section 6 we summarize the findings of this paper while providing some future work that we want to carry out along the same lines of research.

2. Emotional speech corpus

The emotional speech corpus used for this study is a self-recorded database consisting of three pairs of acted emotions uttered by a professional Japanese voice actress: happy - sad, calm - insecure, excited - angry in addition to neutral reading speech. A detailed description of the amounts of data per emotion can be seen in table 1.

For the recordings of the above emotional speech data in a studio booth, the voice actress was instructed to use consistent acoustic realization for each emotion and to maintain emotional strength (rather than changing emotional expressions and

Table 1: *Description of the Japanese emotional speech database. Audio duration includes silences of the beginning and end of the utterance and is expressed in minutes. Speaking rate excludes silences and is expressed in phones per second. Total duration and average speaking rate for the whole database are also shown. Phone alignment was obtained based on HMM-based forced alignment.*

Emotion	#Sentences	Audio duration	Speaking rate
Neutral	1200	147 min	10.39 phones/sec
Happy	1200	133 min	10.90 phones/sec
Sad	1200	158 min	9.04 phones/sec
Calm	1200	154 min	9.05 phones/sec
Insecure	1200	141 min	9.88 phones/sec
Excited	1200	136 min	10.51 phones/sec
Angry	1200	148 min	9.26 phones/sec
Total	8400	1017 min	9.86 phones/sec

Table 2: *Description of the Japanese emotional database recording sentences. The third 'common' column indicates if the sentences were used for recordings of other emotional categories.*

Source	Sentences	Common
News	101	Yes
Novel	313	No
TED talks	196	Yes
Car navigation system	200	Yes
MULTEXT	191	Yes
Phonetically balanced	199	Yes
Total	1200	

strength depending on the meanings of sentences every time) in order to minimize variations within each emotion [14]. This is important because this reduces uncontrollable factors caused by the speaker and allows us to analyze listener's factors easier.

The recorded sentences were chosen to be without emotional meaning, and hence may also be used for recordings of other emotional categories. This choice was made because we aim to analyze misperceptions due to acoustic cues rather linguistic cues. Such sentences were carefully chosen from conversational text resources such as TED talks or MULTEXT [15], rather than from news text resources. Conversational texts were easier for the voice talent to express the emotions compared to news sentences. We have also used novel sentences for the recording, but we manually filtered out sentences that induced emotional context emotion-by-emotion. Phonetically balanced sentences were also recorded so we can build a speech synthesizer from this database. Please see Table 2 for the breakdown.

3. Generating emotional speech under adverse conditions

In order to reproduce realistic emotional speech under noisy and reverberant environments, we have used various types of stereo noise recorded using a binaural head and torso mannequin in a driving car and room-impulse response in the same car. In the following subsections, we explain how we prepared the stimuli for the listening test.

Table 3: *Recorded noise conditions. All kinds of noise were recorded in both kinds of route with the exception of open windows, which was only recorded in the city route.*

Routes	**In-car conditions**
City route (CR)	Closed windows (CW)
Highway route (HR)	Open windows (OW, CR only)
	Competing speaker (CS)

3.1. Recording in-car noises

For recording the in-car noises we have used a B&K 4100 head-and-torso mannequin, which was placed in the front passenger seat of a Toyota Aqua hybrid car. The mannequin was fixed to the seat with the car seat-belts and a B&K WA-1647 car seat fixture. We recorded noise in a number of conditions on city routes and highways near Tokyo, Japan [12]. The recorded noise conditions that have been considered for the present research can be seen in Table 3. Between six minutes and ten minutes of in-car noises were recorded per condition and all material was down-sampled to 48 kHz and high-pass filtered to attenuate noise found below 70 Hz.

On the city routes, the average car speed was 40-60 km/h while the the average car speed on the highways was 80-100 km/h. For the closed windows, we turned on the air conditioner. In the windows open condition, we opened half-way the window closest to the driver. A competing-speaker noise condition was recorded by playing pre-recorded speech material from a male speaker talking in English using a loudspeaker positioned in the backseat of the car. The loudspeaker was placed at a particular height simulating a person sitting in the middle of the back of the car. This condition was only recorded with closed windows. An English male competing-speaker masker was chosen to contrast with the Japanese female speaker of the emotional speech database. Because of the language differences between the masker and target speaker, there is no information (i.e. language) masking.

Finally, we have also recorded the room impulse responses (RIRs) using the front loud-speaker of the same car. For the recording of RIRs, we have parked the car in an indoor garage and, with the windows closed and also with the air-conditioner turned off, we played a sine sweep signal to accurately measure RIRs using FuzzyMeasure [16]. The final RIR was the minimum phase version of those generated by the tool.

3.2. Mixing noise and speech

When you want to artificially produce noisy speech samples using the varying real noises, adding noise at a particular signal-to-noise ratio (SNR) is not such a simple process. We must clearly characterize both speech and noise audio samples and normalize them according to perceptual standards so that the combined version presents the desired SNR. In this research, we considered mainly the variations in noise power, but, we assumed that it made a sense to utilize the averaged speech power within one utterance even if there could be fluctuations.

The processes that we followed to generating the desired noisy speech samples are as follows:

3.2.1. Filter the noise according to "A" standards and intensity calculation

It is known that not all frequencies of noises are perceived in the same fashion as speech, which is why A-weighting standard

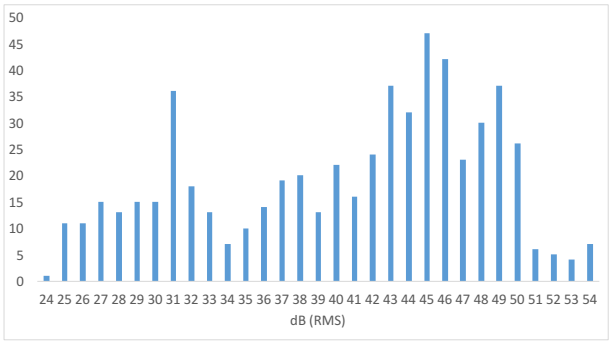

Figure 1: *Histogram of the noise intensities in the city-route open-window condition. The horizontal axis shows the intensity in dB (RMS) of the segmented audio files and vertical axis shows frequency of each bin.*

was introduced [17]. To obtain the perceptual noise profiles, we applied A-weighting to long audio files in the different noise conditions. Then, we split the noise files to segments of the duration of all the target emotional speech samples and computed their intensity in dB (RMS) using Praat's estimator [18].

3.2.2. Characterize the noise profiles

Since we had filtered and spited the audio files now, we could obtain the noise profiles that were to be merged with the audio samples. An example of the obtained noise profiles for the OW condition can be seen in Figure 1. Then, we flagged all the audio splits within the mean $\pm$ one standard deviation as "assignable" samples. This allowed us to consider standard representatives of each noise condition instead of rare noise events.

3.2.3. Characterize target speech intensity and convolve with the RIRs

Next, in order to simulate "listening to a speech utterance played from the front loudspeakers at the front left seat in the car", we convoluted the speech samples with the RIRs recorded by a microphone within left ear or right ear of the head-and-torso mannequin individually. This convolution converted the monaural clean speech samples into stereo speech samples where left and right channel have slightly different degrees of reverberation. Moreover, in order to characterize speech intensity, we used the ITU-T Recommendation P.56 (V56) [19] so as to consider only active speech frames. We then obtained the intensity in dB (RMS) of each emotional speech sample.

3.2.4. Merge noise and speech to generate the target sample

Finally, once we had characterized both noise and speech intensities, we randomly assigned noise samples of the adequate length and from every noise condition to every emotional speech sample and mixed them after amplification and attenuation according to the ratio that we want to achieve. In general we opted to attenuate whenever possible for the desired condition so as not to introduce distortion in the samples and so as to be wary of clipping issues. The last step was to normalize every audio sample to -10 dB under clipping.

4. Perceptual evaluation

We generated emotional speech samples in all the noise conditions using both natural and vocoded speech for all 7 emotions (neutral, happy, sad, excited, angry, calm and insecure).

Table 4: *Age and gender of the evaluators.*

Age	Count
18-29	71
30-39	143
40-49	175
50-59	89
60+	15

Gender	Count
Female	260
Male	154

Table 5: *Results on EIR in percentages, averaged across noise conditions. Nat stands for natural speech and Voc for vocoded speech. The asterisk at -5dB SNR denotes that EIR changes between 0dB and -5dB are statistically significant according to student's t-distribution.*

Emotion	Nat 0dB	Nat -5dB	Voc 0dB	Voc -5dB
Neutral	74.7	74.5	73.3	*77.4
Positive				
Happy	84.2	83.2	82.6	83.7
Calm	68.4	*60.4	64.7	*53.4
Excited	33.4	34.7	29.3	30.1
Negative				
Sad	83.8	82.9	81.6	*79.9
Insecure	74.6	*72.0	69.1	*66.6
Angry	89.7	*86.9	86.2	*84.3
Average	72.7	*70.7	69.5	*68.1

Concretely, we evaluated the 8400 natural speech samples of the database. Further we also evaluated 100 vocoded speech samples per emotion randomly selected from the corpus. The vocoded speech was obtained by analysis-by-synthesis using the WORLD vocoder [20]. In total all the natural and vocoded samples were evaluated once in each noise condition for two different SNR values: 0dB and -5dB.

The evaluation was carried out by crowd-sourced Japanese native speakers. The evaluators were presented a web-page where they first had to input their gender and age. Then they were asked to rate 14 utterances in the above noise conditions. They were able to play the samples as many times as they wanted. For the emotional recognition question, they were asked to select an answer between a pool of 9 emotions: neutral, happy, sad, excited, angry, calm, insecure, surprised, bored and "other". For the perceived emotional strength they were asked to rank in the MOS scale: from "1 - almost no emotion" to "5 - very emotional". They were also allowed to answer "6 - no emotion". The 14 utterances were selected so that every emotion was rated twice. Whether the sample was natural or vocoded speech was assigned randomly. Evaluators were allowed to repeat the task up to 20 times to reduce the number of required evaluators but without allowing them to repeat the task so many times. A total of 414 people took part in the evaluation, for a total of 91,000 data points. Age and gender distributions of the evaluators are shown in Table 4

5. Evaluation results

5.1. Results in terms of noise conditions and signal-to-noise ratio

Table 5 shows the correctly-identified ratio of emotions, emotion-identification-rates (EIR) across noise conditions. In the table, the asterisks at -5dB SNR denote that EIR changes between 0dB and -5dB are statistically significant according to student's t-distribution. By looking averaged EIRs across all

Table 6: *Results on perceived ES scores averaged across conditions. The asterisk at -5dB SNR denotes that the ES score changes between 0dB and -5dB are statistically significant according to student's t-distribution.*

Emotion	Nat 0dB	Nat -5dB	Voc 0dB	Voc -5dB
Neutral	2.63	*2.53	2.51	2.52
Positive				
Happy	3.59	*3.51	3.56	*3.41
Calm	3.04	*2.74	2.96	*2.66
Excited	3.41	*3.32	3.38	*3.23
Negative				
Sad	4.08	*3.94	4.08	*3.87
Insecure	3.27	*3.00	3.19	*2.80
Angry	3.78	*3.56	3.68	*3.40
Average	3.49	*3.32	3.43	*3.22

Table 7: *Comparison of the noise conditions averaged across emotions. EIRs shows percentages and ES shows MOS scores.*

Route	Nat 0dB	Nat -5dB	Voc 0dB	Voc -5dB
EIR				
CR	72.8	*70.4	70.7	*67.9
HR	72.5	*71.0	67.6	68.4
ES				
CR	3.50	*3.32	3.46	*3.22
HR	3.48	*3.33	3.38	*3.22

the emotions, we observed a decrease of 2% (a 7.4% relative increase in error rate) in EIR for natural speech between the 0dB and the -5dB SNR condition. Looking at the effect of noise in vocoded speech, we see a decrease of 1.4% in EIR (6.8% relative increase), very similar in relative rates when compared to natural speech. Most of the differences between 0dB and -5dB SNRs are statistically significant. As we hypothesized, not only speech intelligibility but also interpretation of the emotional content of an utterance may become worse when listening to natural or synthetic speech in adverse conditions.

Table 6 shows the perceived emotional strength (ES) scores averaged across noise conditions. Again in the table, the asterisks at -5dB SNR denote that ES changes between 0dB and -5dB are statistically significant. Looking at the results of natural speech, we can see that all the differences between 0dB and -5dB SNRs are statistically significant and on average there is a decrease of 0.17 points. We can observe the same tendency for the vocoder results (average decrease of 0.21).

Table 7 shows a comparison of EIR and ES with respect to the route conditions. From this table, we can see that the highway route conditions slightly present lower EIRs and ES at 0dB compared to the city-route conditions probably due to higher speed of the car. But the more obvious changes in both aspects of emotion perception were caused by the SNR levels.

5.2. Results in terms of emotional categories

Next we analyze EIRs in Table 5 per emotional category. It clearly shows that one of the positive emotion named "calm" have the most severe degradation in terms of EIR, with a decrease of 8% (a 25% relative increase in error), while there is no significant degradation in the recognition rates for neutral, happy or sad voices. The rest of the emotions show results a small but significant degradation of 2%. The vocoded emotional speech has the same trend. The calm voice has a significant decrease in recognition rates of 11.3% when we compare 0dB with -5dB. Finally if we compare the EIR changes of positive and negative emotions overall, we see that the EIR changes of three negative emotions are statistically significant while that of only one positive emotions are statistically significant in our experiment. This implies that the positive and high arousal emotions of our voice actress seem to be more robust to in-car noise compared to her negative and low arousal emotions.

5.3. Results in terms of listener's factors

To look at the effect of gender and age, we carried out multivariate and uni-variate ANOVA analyses. In order to do this, we split the evaluators in 5 age groups shown in Table 4. The analysis reveals significant effects below.

- Gender plays a role in both EIR and ES: in general our female listeners showed significantly higher recognition capabilities (1% higher EIR) and perceived emotions as stronger (0.1 larger ES scores).
- Age plays an even bigger role in both EIR and ES: younger people seem to be better at recognizing emotions under adverse conditions (4% higher EIR) and tend to perceive them as significantly stronger (0.3 larger ES scores).
- A combination of age and gender shows a very considerable variability: younger females surprisingly showed in average 12% better EIR and 0.3 higher ES compared to older males.

6. Conclusions and future work

In this research, beginning with our emotional speech database, we have recorded a number of real noise samples inside a car and the room impulse response from the front loudspeakers with a stereo microphone placed in the front seat of the car. Then we have generated emotional speech samples under the adverse conditions for 0dB and -5dB SNRs according to a strict procedure to mimic as closely as possible the effect of listening to said emotional samples in a running car. Finally we carried out a massive crowd-sourced evaluation where 414 people evaluated a total of 91,000 samples.

The evaluation has verified our hypothesis that noise plays a significant role in emotion perception, that is, significantly reducing listener's capabilities of recognizing them as conditions grow worse. Not only that, but we have also seen how listeners tend to perceive emotions as being less strong in noisier environments. These effects become more noticeable when talking about vocoded speech, which strongly suggests that it is necessary to take this effect into account when thinking about including emotional synthetic speech in applications to be used in noisy environments (e.g. car navigation systems). We also found that gender and age of listeners also influenced the results.

Our planned future work mainly consists in designing a method to automatically take into account both environmental conditions and target listener for generating highly customized emotional speech in order to maximize the expressive capabilities of our emotional text-to-speech system.

Acknowledgements: The work presented in this paper was partially supported by TOYOTA motor corporation.

7. References

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[2] S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, J. Sotelo, A. Courville, and Y. Bengio, "Samplernn: An unconditional end-to-end neural audio generation model," *arXiv preprint arXiv:1612.07837*, 2016.

[3] S. O. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, J. Raiman, S. Sengupta *et al.*, "Deep voice: Real-time neural text-to-speech," *arXiv preprint arXiv:1702.07825*, 2017.

[4] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfiltering for statistical parametric speech synthesis," in *ICASSP*, 2017.

[5] H.-T. Luong, S. Takaki, G. E. Henter, and J. Yamagishi, "Adapting and controlling DNN-based speech synthesis using input codes," in *ICASSP*, 2017.

[6] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," 2016.

[7] Y. Tang, M. Cooke *et al.*, "Glimpse-based metrics for predicting speech intelligibility in additive noise conditions," in *Proceedings of the Annual Conference of the International Speech Communication Association*. International Speech and Communication Association, 2016, pp. 2488–2492.

[8] P. N. Petkov and Y. Stylianou, "Adaptive gain control for enhanced speech intelligibility under reverberation," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1434–1438, 2016.

[9] D. Fogerty, J. B. Ahlstrom, W. J. Bologna, and J. R. Dubno, "Sentence intelligibility during segmental interruption and masking by speech-modulated noise: Effects of age and hearing loss," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. 3487–3501, 2015.

[10] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Synthesis and perception of breathy, normal, and lombard speech in the presence of noise," *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, 2014.

[11] E. Jokinen, M. Takanen, M. Vainio, and P. Alku, "An adaptive post-filtering method producing an artificial lombard-like effect for intelligibility enhancement of narrowband telephone speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 619–628, 2014.

[12] C. Valentini-Botinhao and J. Yamagishi, "Speech intelligibility in cars: the effect of speaking style, noise and listener age," in *Submitted to Interspeech*, 2017.

[13] J. H. Rindel and C. L. Christensen, "Room acoustic simulation and auralization–how close can we get to the real room," in *Proc. 8th Western Pacific Acoustics Conference, Melbourne*, 2003.

[14] A. Athanasopoulou and I. Vogel, "Acquisition of prosody: The role of variability," *Speech Prosody 2016*, pp. 716–720, 2016.

[15] K. Shigeyoshi, K. Tatsuya, M. Kazuya, and I. Toshihiko, "Preliminary study of japanese multext: a prosodic corpus," in *International Conference on Speech Processing, Taejon, Korea*, 2001, pp. 825–828.

[16] SuperMegaUltraGroovy. Fuzzmeasure. [Online]. Available: http://supermegaultragroovy.com/products/fuzzmeasure/

[17] R. L. S. Pierre Jr, R. Acoustics, D. J. Maguire, and C. S. Automotive, "The impact of A-weighting sound pressure level measurements during the evaluation of noise exposure," in *Conference NOISE-CON*, 2004, pp. 12–14.

[18] P. P. G. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[19] I. Ree, "P. 56: Objective measurement of active speech level," 1993.

[20] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.