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

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 recently made remarkable progress, and we now have the technology to generate natural-sounding synthetic speech almost comparable to human speech. In particular, approaches based on convolutional neural networks [1], waveform modeling [2, 3] and generative-adversarial networks [4] have proven extremely successful in generating high-quality synthetic speech (and other types of sounds, such as onomatopoeia or music). Our group has 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]. Other studies have proven being 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. Speech intelligibility is known to deteriorate significantly under noisy conditions [7, 8], and this effect becomes even more significant with age [9]. Human verbal communication implicitly compensates for this degradation through Lombard speech [10]. There have been several attempts to use environmental noise information to adaptively improve the intelligibility of speech synthesizers [11, 12].

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

We performed a large-scale listening test in which over 400 crowdsourced listeners between the ages of 21 and 72 assessed natural and vocoded acted emotional speech stimuli. In emotional speech synthesis, acted emotions are typically used for acoustic modelling. Hence, in our experiment we used 7 speech uttered by a professional voice actress expressing seven different emotions. Vocoded speech was used as a proxy for text-to-speech systems, which show slightly worse quality than natural speech. Both types of stimuli were then artificially degraded using the room impulse response (RIR) recorded in a car [13] and various in-car noise types recorded in the same car . The listeners were asked to identify the emotional content of the resulting speech utterances under adverse conditions and to rate the perceived strength of the emotional content. We then analyzed the resulting emotional recognition rates and perceived emotional strength with respect to the signal-to-noise ratios (SNRs), emotional categories, and listener' age and gender.

The paper is structured as follows. Section 2 introduces the emotional speech corpus that we used for the evaluation, and section 3 explains how we recorded the noise added to the emotional speech and the exact procedure by which it was added. In section 4 we describe the evaluation process, before discussing the results in section 5. Finally in section 6 we summarize the findings of this paper while indicating future work that we want to perform along the same lines of research.

2. Emotional speech corpus

The emotional speech corpus used for this study was a self-recorded database consisting of three pairs of acted emotions uttered by a professional Japanese voice actress: happy sad, calm insecure, and excited angry in addition to neutral reading speech. Table 1 gives a detailed breakdown of the amount of data for each emotion.

For recording of the emotional speech data in a studio booth, the voice actress was instructed to use a consistent

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-532

Table 1: *Summary of the Japanese emotional speech database. Duration includes silences at the beginning and end of an 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 listed. Phone alignment was obtained through HMM-based forced alignment.*

Emotion	#Sentences	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: *Breakdown of the sentences recorded for the Japanese emotional speech database. The third column, labeled "Common" indicates whether the sentences in that category were also used for 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
All	1200	

acoustic realization for each emotion and to maintain emotional strength (rather than changing the emotional expressions and strength according to the meaning of the sentence each time), in order to minimize variation within each emotion [14]. This was important because it reduced uncontrollable effects caused by the speaker and allowed us to more easily analyze the listeners' behaviors.

The recorded sentences were chosen to be without emotional meaning, so that they could be used for recordings of multiple emotional categories. We chose this approach because we aimed to analyze misperceptions due to acoustic cues rather than linguistic cues. Such sentences were carefully chosen from conversational text resources, such as TED talks or MULTEXT [15], rather than only from news text resources. Conversational texts made it easier for the voice actress to express emotions, as compared to news texts. We also used sentences from novels, but we manually removed sentences that induced a specific emotional context, emotion-by-emotion. Phonetically balanced sentences were also recorded so that we can build a speech synthesizer from this database. Table 2 gives a breakdown of the recorded sentences.

3. Generating emotional speech under adverse conditions

To reproduce realistic emotional speech in noisy, reverberant environments, we used various types of stereo noise recorded by using a binaural head and torso mannequin in a driven car and the room-impulse response (RIR) 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 was recorded on both kinds of routes, with the exception of noise from open windows, which was only recorded on the city route.*

Route types	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 noise, we 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 attached to the seat with the car's seat belt and a B&K WA-1647 car seat fixture. We recorded noise under a number of conditions on both city and highway routes near Tokyo, Japan [12]. Table 3 lists the recorded noise conditions that were used. Between six and ten minutes of in-car noise was recorded under each condition, and all material was down-sampled to 48 kHz and high-pass filtered to attenuate noise found below 70 Hz.

On the city route, the car's range of speed was between 40 and 60 km/h, while the range on highways was 80-100 km/h. With closed windows, we turned on the air conditioner. In the open-windows condition, we opened the window closest to the driver halfway. A noise condition with a competing speaker was recorded by playing pre-recorded speech from a man talking in English through a loudspeaker placed in the back seat of the car. The loudspeaker was placed at a particular height to simulate a person sitting in the middle of the back seat. This condition was only recorded with closed windows. An English male competing-speaker masker was chosen to contrast with the Japanese female speaker in the emotional speech database. Because of the language difference between the masker and the target speaker, there is no information (i.e. language) masking.

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

3.2. Mixing noise and speech

In artificially producing noisy speech samples by using a variety of real noises, adding noise at a particular SNR is not a simple process. Both speech and noise audio samples must be clearly characterized and then normalized according to perceptual standards so that the combined signal exhibits the desired SNR. In this research, we mainly considered variations in noise power, but we also assumed that it made sense to use the average speech power within one utterance even if there could have been fluctuations.

We applied following steps to generate the desired noisy speech samples:

3.2.1. Filter noise according to A-weighting and intensity calculation

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

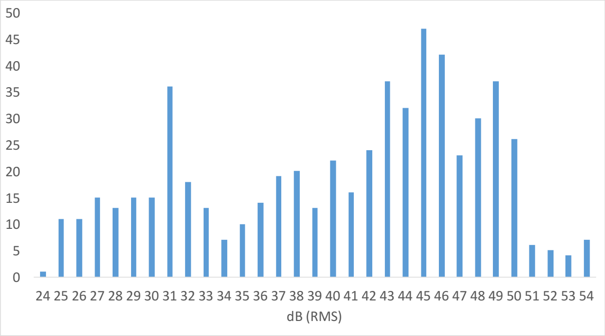

Figure 1: *Histogram of noise intensity under the city-route open-window condition. The horizontal axis indicates the RMS intensity in dB of the segmented audio files, and the vertical axis indicates the frequency of each bin.*

was introduced [17]. To obtain perceptual noise profiles, we applied A-weighting to long audio files under the different noise conditions. Then, we split the noise files into segments of the same durations as all of the target emotional speech samples and computed their RMS intensities in dB by using the estimator in Praat [18].

3.2.2. Characterize noise profiles

Since we had now filtered and split the audio files, we could obtain the noise profiles to be merged with the audio samples. Figure 1 shows an example of the noise profile obtained under the OW condition. Then, we flagged all the audio splits within the mean ± 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 RIRs

Next, to simulate the experience of listening to a speech utterance played from the front loudspeakers while sitting in the front left seat of the car, we recorded the RIR using a microphone within the left and right ear of the head-and-torso mannequin, then we convolved the speech samples with each ear response. This convolution converted the monaural clean speech samples into stereo speech samples in which the left and right channels had slightly different degrees of reverberation. Moreover, to characterize the speech intensity, we applied ITU-T Recommendation P.56 (V56) [19] so as to consider only active speech frames. We then obtained the RMS intensity in dB of each emotional speech sample.

3.2.4. Merge noise and speech to generate the target sample

Finally, once we had characterized both the noise and speech intensities, we randomly assigned noise samples of adequate length and from every noise condition to every emotional speech sample. We then mixed the combined samples after amplification and attenuation to obtain the desired SNR. In general, we opted to attenuate whenever possible for the desired condition, so as to avoid introducing distortion in the samples and to be wary of clipping issues. The final step was to normalize every audio sample to -10 dB below clipping.

Table 4: *Age and gender distributions 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: *Emotion identification rates (EIRs) in percentages, averaged across noise conditions. ”Nat” stands for natural speech and ”Voc” for vocoded speech. The asterisks indicate EIR changes between 0 and -5 dB that were statistically significant according to Student’s t-tests.*

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

4. Perceptual evaluation

We generated emotional speech samples under all noise conditions by using both natural and vocoded speech for all seven emotional conditions (neutral, happy, sad, calm, insecure, excited and angry). We then concretely evaluated the 8400 natural speech samples in the database. For each emotion we also evaluated 100 vocoded speech samples randomly selected from the corpus. The vocoded speech was obtained through analysis-by-synthesis with the WORLD vocoder [20]. All the natural and vocoded samples were evaluated once under each noise condition for two different SNR values: 0 dB and -5 dB.

The evaluation was performed by crowdsourced Japanese native speakers. The evaluators were first shown a web page on which they had to input their gender and age. They were then each asked to rate 14 utterances under the above noise conditions. They were able to play each sample as many times as they wanted. For emotional recognition, they were asked to select an answer from a pool of 10 emotions: neutral, happy, sad, calm, insecure, excited, angry, surprised, bored and other. For the perceived emotional strength they were asked to rank each utterance 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 evaluated twice. Each sample was randomly selected from either natural or vocoded speech. To reduce the number of evaluators required, each evaluator was allowed to repeat the task up to 20 times. A total of 414 people took part in the crowdsourced evaluation, for a total of 91,000 data points. Table 4 lists the age and gender distributions of the evaluators.

5. Evaluation results

5.1. Results in terms of noise condition and SNR

Table 5 lists the rates of correctly identified emotions across noise conditions. We refer to this rate as the emotion identification rate (EIR). In the table, the asterisks indicate EIR changes

Table 6: *Perceived ES scores averaged across noise conditions. The asterisk at -5dB SNR indicates ES score changes between 0 and -5 dB that were statistically significant according to Student's t-tests.*

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 noise conditions averaged across emotions. EIRs expressed as percentages, and ES, as MOS.*

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

between 0 and -5 dB that were statistically significant according to Student's t-tests. By considering averaged EIRs across all emotions, we observed a decrease of 2% (a 7.4% relative increase in error rate) in EIR for natural speech between the SNRS of 0dB and the -5dB SNR. For vocoded speech, we observed a decrease of 1.4% in EIR (6.8% relative increase), very similar in relative rate as compared to natural speech. Most of the differences between the SNRS of 0 and -5 dB were 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 under adverse conditions.

Table 6 lists the perceived ES scores averaged across noise conditions. The asterisks again indicate changes between 0 and -5 dB that were statistically significant. Considering the results for natural speech, we can see that all the differences between SNRs were statistically significant, with an average a decrease of 0.17 points. We can observe the same tendency for the vocoded results (average decrease of 0.21).

Table 7 compares both the EIR and ES with respect to the different route conditions. We can see that the highway route condition exhibited slightly lower EIR and ES at 0dB as compared to the city route conditions, probably because of the changes in noise spectral content due to the car's higher speed. The more obvious changes in both aspects of emotional perception, however, were caused by the SNR levels.

5.2. Results in terms of emotional categories

Next, we analyze the EIRs listed in Table 5 in terms of emotional category. The results clearly show that the positive emotion "calm" suffered the greatest degradation in terms of EIR, with a decrease of 8% (a 25% relative increase in error). In contrast, there was no significant degradation in the recognition rates for neutral, happy, or sad voices. The rest of the emotions exhibited a small but significant degradation of 2%. The vocoded emotional speech samples showed the same trend. The calm voice had a significant decrease in EIR of 11.3% from 0 dB to -5 dB. Finally, if we compare the results for the positive and negative emotions overall, we can see that the EIR changes of all three negative emotions were statistically significant, while only one positive emotion was for our experiment. This implies that positive and high-arousal emotional speech from the voice actress was more robust with respect to in-car noise, compared to her negative and low-arousal emotional speech.

5.3. Results in terms of listener's age and gender

To consider the effects of gender and age, we carried out both multivariate and univariate ANOVAs. To do this, we split the evaluators into the five age groups listed in Table 4. The analyses revealed the significant effects below.

- Gender played a role in both EIR and ES: in general, our female listeners exhivited significantly higher recognition capability (1% higher EIR) and perceived emotions as stronger (0.1 greater ES scores).
- Age played an even bigger role in both EIR and ES: younger people seemed better at recognizing emotions under adverse conditions (4% higher EIR) and tended to perceive them as significantly stronger (0.3 greater ES scores).
- Combinations of age and gender showed considerable variability: younger women surprisingly exhibited 12% better EIR and 0.3 higher ES, on average, as compared to older men.

6. Conclusions and future work

In this research, we recorded an emotional speech database, a number of real noise samples inside a car and the RIR from the front loudspeakers with a stereo microphone placed in the front left seat. Then, we generated emotional speech samples under adverse conditions with SNRS of 0 dB and -5 dB according to a strict procedure designed 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 in which 414 people evaluated a total of 91,000 samples.

The evaluation verified our hypothesis that noise plays a significant role in emotional perception, in particular, by significantly reducing a listener's capability for recognizing emotions as conditions grow worse. Moreover, we also observed that listeners tended to perceive emotions as weaker in noisier environments. These effects became more noticeable for vocoded speech. This strongly suggests the importance of accounting for this effect when thinking about including emotional synthetic speech in applications meant for noisy environments (e.g., car navigation systems). We also found that the gender and age of listeners also influenced the results.

Our planned future work mainly consists of designing a method to automatically consider both environmental conditions and the target listener in generating highly customized emotional speech, in order to maximize the expressive capabilities of emotional text-to-speech systems.

Acknowledgement: 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.