ISASE 2019

Modelling of Non-Linguistic Utterances for Machine to Human Communication in Dialogue

Ahmed KHOTA*, Asako KIMURA ** and Eric COOPER ***

Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, 525-8577, Japan
** gr0341xp@ed.ritsumei.ac.jp*
*** asa@rm.is.ritsumei.ac.jp*
**** cooper@is.ritsumei.ac.jp*

Abstract: Non-Linguistic Utterances (NLUs) present a potentially useful alternative communication channel between humans and machines. NLUs are potentially cheaper, and easier to implement, and not limited to the constraints of natural language and therefore may be appropriate in situations such as assisting tourists with various language backgrounds and needs. An experiment was done to establish ranges for NLU parameters such as pitch, duration, amplitude, and timbre. Subjects listened to randomly produced NLUs and selected applicable dialogue descriptors from: Positive, Negative, Greeting, Apology, Thanking, Hesitation, Question, Approval, Disapproval, Hushing, None of the Above. Factor analysis yielded 3 major factors, which were labeled as follows: 1) Affirmative vs Negative, 2) Questioning, and 3) Meaningful vs Indeterminate. NLUs with lower pitches, downward pitch patterns, and simpler timbres were found to be more Negative. Those with upward pitch patterns were more likely to be identified as a Question. In future work, experiments will be used to develop a model of NLU inference and interpretation within a Dialogue in terms of the dominant descriptors and to test the model in applications for tourist support systems.

Keywords: *Non-Linguistic Utterances, NLU, Human-Robot Interaction, Dialogue,*

1. INTRODUCTION AND PREVIOUS WORK

Non-Linguistic Utterances (NLUs) for communication from machine to human in a dialogue setting have been popularized in fiction, for example in the Star Wars movies where the robot R2D2 communicates with his human counterparts using squeaks, beeps, and other robotic sounds. NLUs are also already used in daily life, for example in train stations to indicate passing a turnstile or an approaching train.

NLUs fall under the umbrella term Semantic Free Utterances (SFUs) which also includes Gibberish Speech, Musical Utterances, and Paralinguistic Utterances. NLUs are sounds that contain no discernible words, are not specifically musical, and exclude laughing or onomatopoeia. They are used to convey information, affect, or to communicate. Their acoustic parameters can be derived from their natural language or real-world analogues [1].

NLUs and other SFUs have been successfully interpreted in terms of affect and emotional expression [2-5]. In most cases, the alteration of the pitch of the sound has had the biggest influence on the way it was interpreted. [1]

Previous research has investigated whether NLUs can successfully convey emotion or affect. Usually, the embodied agent is a robot of some sort. Reasons for using NLUs include the following. Natural language programming is costly and difficult [1,5]. Not all applications necessarily require advanced natural language communication [1]. Programming for multiple languages adds additional complexity [1,4,5]. NLUs provide two main benefits: they are not linked to any language, and they can communicate a message in a very short time [6].

Fernandez De Gorostiza et al. [6] proposed that the NLUs can be used as complimentary in any communication system to enhance expressiveness, eloquence, and efficiency of interactions. Users naturally expect more capability from systems using natural languages, and therefore NLUs may reduce user expectations according to actual capabilities [1,6].

Fernandez De Gorostiza et al. developed a method for the generation of NLUs for their Sonic Expression System. The central concept that they created was the idea of the quason, which they define as "the smallest sound unit that holds a set of indivisible psychoacoustic features that makes it perfectly distinguishable from other sounds, and whose combinations generate a more complex individual sound unit” [6]. The major characteristics of the quason are its amplitude, frequency, and time. Individual quasons combine to form the sonic expression. The researchers developed sonic expressions for the following types of communicative acts: Agreement, Hesitation, Denial, Questioning, Hush, Summon, Encouragement, Greeting, Laughter. Each sound was designed at three levels of intensity and evaluated by 51 participants [6]. The method of using quasons, and the analysis of their frequency, amplitude and time, is applicable to the current research in terms of modelling interpretable NLUs within a dialogue.

2. OBJECTIVES AND CONTEXT

The objective of the current research is to develop, test, and validate a model for using NLUs in Human Machine Communication in a dialogue setting.

The context for the practical validation of the model will be a tourist assistance setting. For example, in many popular tourist destinations, tourists are often lost or searching for guidance along their route. In such situations, NLUs may provide a supplementary option to the existing information channels. Advantages of NLUs in this application include that they are not bound by any human spoken language, don't interfere with surrounding communication, may be small and inconspicuous, and may enhance user experience.

3. EXPERIMENTAL METHODS

3.1 Purpose and methodology

The purpose of the experiment described below is to gather data on the interpretation of randomly produced NLU sounds and use the data to narrow down the range of sounds to those that are interpreted as dialogue. The NLU sounds are random in terms of the number of and value of pitch changes, timbre (sine, sawtooth, square wave combination), amplitude (attack, decay, sustain, release), and duration. Figure 1 shows how the sounds were generated (using Pure Data graphical programming). Pitches between 31 (49Hz) and 91 (1568Hz) were used as these were found in testing to fall inside a comfortable range, i.e. not too low and not too high pitched to listen to comfortably. Quason times of between 200ms and 1000ms were used to make up the NLUs (each NLU contains multiple quasons strung together end-to-end), which themselves were limited to a maximum of 2500ms. It was found in testing that sounds longer than 2000ms were generally harder to interpret. The following ranges were used: Attack time (0-500ms), Decay time (0-500ms), Sustain value (0-1), and Release time (0-500ms). An implementation of the Glide (or Portamento) concept from music theory (the smooth transition from one pitch to the next) was used to apply random levels of smoothing to the transitions between the quasons.

In the experiment, the random NLU sounds are played for the subjects. 28 audio clips in total are played one after the other, and the subjects listen through headphones. After listening to each clip, subjects select all applicable descriptors for each sound from the following list, taken from combinations of similar definitions from previous work [1-6]: Positive, Negative, Greeting, Apology, Thanking, Hesitation, Questioning, Approval, Disapproval, Hushing, None of the Above. Figure 2 shows the process described.

3.2 Results

10 subjects participated in the experiment, 2 females, 8 males, aged 25-30, from 7 nationalities. The data per NLU is shown in Table 1.

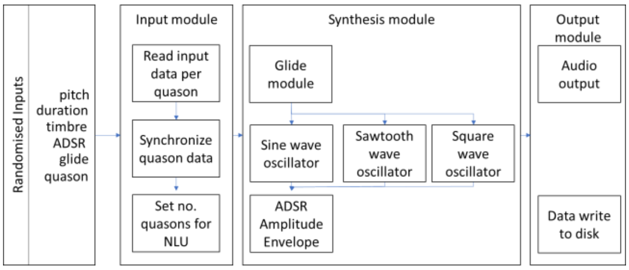

Figure 1: Sound Design

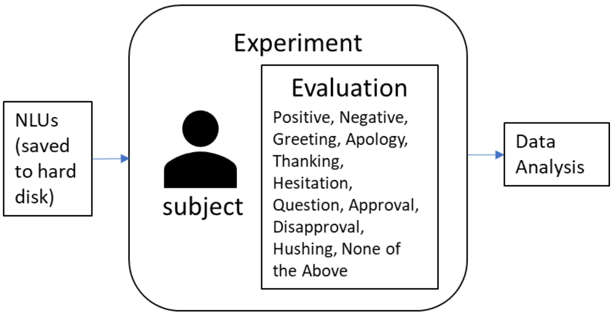
Experiment
NLUs (saved to hard disk)
subject
Evaluation
Positive, Negative, Greeting, Apology, Thanking, Hesitation, Question, Approval, Disapproval, Hushing, None of the Above
Data Analysis

Figure 2: Experimental Methodology

The timbre score is the total number of active oscillators (sine, sawtooth, square), for a maximum of 3. The timbre score for each NLU is the total timbre score for its quasons. E.g. if the NLU contains 3 quasons, the maximum possible timbre score is 9, and minimum 0. The logic is that the higher the timbre score, the more oscillators are active at the same time and therefore the more complex the timbre of each quason and therefore also the overall NLU.

Table 1: Audioclip attributes with Factor Analysis scores

No.	pitches	time (ms)	timbre	D1	D2	D3
1	58, 68, 78	1783	4	-0.96	2.00	0.98
2	54, 33	609	6	-2.23	0.23	0.77
3	61, 79, 55	1327	4	0.61	-0.80	0.31
4	48, 77	1042	4	-0.49	1.53	-0.68
5	80, 67, 73	1915	6	-0.03	0.95	0.83
6	56, 79	1818	3	-0.02	0.24	-0.17
7	32, 35	753	3	-1.68	-1.54	0.51
8	58, 68	1052	4	0.95	1.50	-1.37
9	78, 54, 33	1574	6	-0.15	-0.65	0.03
10	48, 88, 62	939	9	0.90	-0.12	1.37
11	42, 82	1529	1	0.54	-0.46	-1.39
12	36, 80, 35	1761	6	-0.55	-1.44	-0.11
13	81, 63	1710	5	0.03	0.18	-0.92
14	91, 88	1615	4	1.22	-0.67	1.21
15	62, 91, 90, 48	2075	7	-0.06	-0.52	-0.68
16	88, 62, 39	1427	2	-1.07	0.25	0.96
17	48, 77, 35	624	7	0.33	1.23	-0.32
18	56, 34	1003	3	-1.33	-0.66	-0.18
19	75, 80, 60	1863	8	1.74	-0.83	1.90
20	50, 80, 63, 37	1812	8	0.18	-0.57	-1.28
21	91, 49, 43	1857	6	0.27	0.74	-0.96
22	86, 85, 52	1097	2	0.34	1.57	0.73
23	32, 50	830	6	-0.98	-1.13	-0.41
24	56, 79	708	4	0.95	0.65	-0.58
25	57, 61, 77	984	6	1.68	-0.14	0.67
26	76, 77	408	5	0.98	-1.32	-1.18
27	76, 54	1136	1	-0.44	0.05	-0.24
28	88, 62, 39	984	6	-0.73	-0.30	0.20

The response results were checked for internal consistency by calculating Cronbach's Alpha, which was found to be 0.71, indicating a strong level of internal consistency and that in general the opinions of the group of subjects did not contradict one another. The Chi square test statistic was also computed and found to be 0.016, indicating that there was a link between the variables for descriptors and audio clips. Cramer's V was 0.282, indicating a medium magnitude effect size of the data.

The response results did not show any obvious trends for this set of randomly generated sounds. Factor Analysis was done to examine the relation between descriptors and yielded the following three factors; D1) Affirmative vs Negative, D2) Questioning, and D3) Meaningful vs Indeterminate. These three factors were made up of the original descriptors shown in Table 2, with the values obtained after varimax rotation. The Factor scores were compared to Average pitch, Duration, number of quasons, pitch pattern, and timbre.

Table 2: Factor Analysis Results

	D1	D2	D3
Negative	**-0.901**	-0.190	0.153
Positive	**0.690**	0.056	0.448
Greeting	**0.593**	0.023	0.488
Disapproval	**-0.592**	-0.337	-0.018
Hesitation	**-0.589**	0.172	-0.043
Thanking	**0.570**	-0.292	0.026
Approval	**0.529**	0.149	0.222
Hushing	**-0.357**	0.174	0.163
Questioning	0.111	**0.959**	0.006
None of the above	0.100	0.037	**-0.760**
Apology	-0.070	-0.186	**-0.441**

Figure 3 shows the correlation of average pitch of each NLU to the first factor, indicating that as the average pitch increases, so does the factor score, suggesting that NLUs containing generally lower pitches were perceived to be more Negative (r = 0.67). Figure 4 shows that NLUs with the downward pitch pattern (no. of audio clips for each pattern shown above each box) were perceived to be more Negative. Figure 5 (no. of audio clips for each pattern shown above each box) shows that NLUs with the upward pitch pattern were generally perceived to be Questioning. Figure 6 (no. of audio clips for each timbre shown above each box) shows that NLUs with simpler timbres (indicated by a lower value on the x axis) were generally perceived to be more Negative.

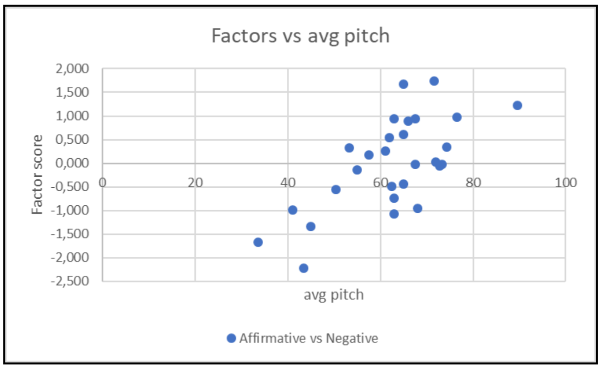

Figure 3: Factor score vs average pitch

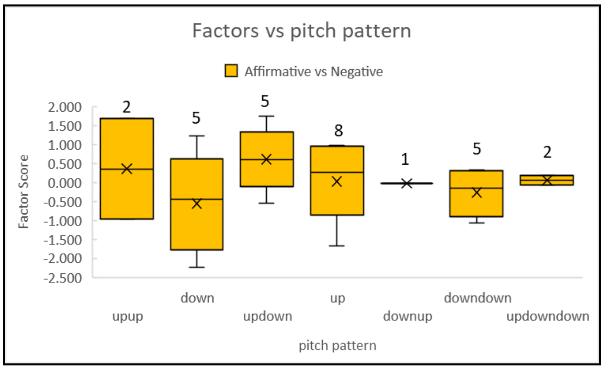

Figure 4: Affirmative vs Negative factor score vs pitch pattern

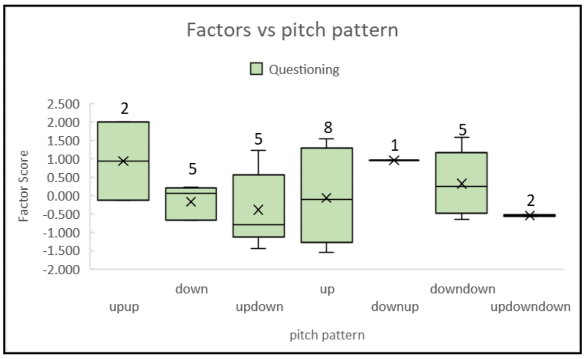

Figure 5: Questioning factor score vs pitch pattern

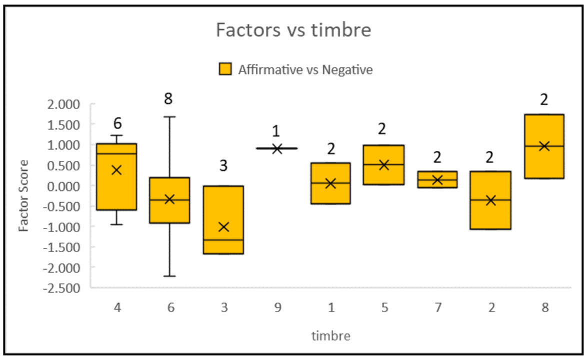

Figure 6: Affirmative vs Negative factor score vs timbre

4. CONCLUSIONS AND FUTURE WORK

An experiment was conducted to establish parameter ranges for NLUs in terms of their interpretation. 10 subjects participated in an experiment, where 28 audio clips were played for them and they selected applicable descriptors from a list that they felt applied to each sound.

Cronbach’s Alpha was calculated to be 0.71, showing a high level of internal consistency in the data. Chi square was 0.016, and Cramer’s V 0.282.

Factor Analysis was done to examine the relation between descriptors and yielded three factors; 1) Affirmative vs Negative, 2) Questioning, and 3) Meaningful vs Indeterminate. It was found that NLUs containing generally lower pitches were perceived to be more Negative. It was found that NLUs with the downward pitch pattern were perceived to be more Negative, while NLUs with the upward pitch pattern were generally perceived to be Questioning. NLUs with simpler timbres were generally perceived to be more Negative. The lack of other clear indicators suggests that NLUs with long durations, high timbre complexity, and many quasons, may have been harder to interpret clearly.

Future work will be focused on new experiments designed to develop the model for interpreting NLUs. The experiments will also use Dialogue parts, where a conversation between two agents, one making NLUs and the other using natural language, will be used. Using the results from this first experiment, the NLUs will either be designed to convey specific affect and/or dialogue parts or be produced randomly but within the ranges stipulated by these experimental results. Mel Frequency Cepstrum [7] will be considered as a tool for audio analysis with a view to defining relevant feature vectors, a method that has been used primarily in speech recognition. A model that would predict NLU interpretation in terms of dominant factor, according to features specified by the Mel Frequency Cepstral Coefficients (MFCC) and the input parameters of each quason (pitch, timbre, duration, amplitude envelope, etc.) would provide a method for generating NLUs for specific applications such as tourist assistance.

REFERENCES

[1] S. Yilmazyildiz, R. Read, T. Belpaeme, W. Verhelst, Review of Semantic-Free Utterances in Social Human–Robot Interaction, International Journal of Human-Computer Interaction, 32(1), 63–85, 2016.

[2] T. Belpaeme, P. Baxter, J. de Greeff, J. Kennedy, R. Read, R. Looije, M. Neerincx, I. Baroni, M. Coti Zelati, Child-robot interaction: Perspectives and challenges. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8239 LNAI, 452–459, 2013.

[3] R. Read, T. Belpaeme, Interpreting non-linguistic utterances by robots, Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments - AFFINE ’10, 65, 2010.

[4] R. Read, T. Belpaeme, How to use non-linguistic utterances to convey emotion in child-robot interaction, Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI ’12, 219, 2012.

[5] R. Read, T. Belpaeme, People Interpret Robotic Non-linguistic Utterances Categorically, International Journal of Social Robotics, 8(1), 31–50, 2016.

[6] J.F. De Gorostiza Luengo, F.A. Martin, A. Castro-Gonzalez, M.A. Salichs, Sound synthesis for communicating nonverbal expressive cues, IEEE Access, 5, 1941–1957, 2017.

[7] M.A. Hossan, S. Memon, M.A. Gregory, A novel approach for MFCC feature extraction, 4th International Conference on Signal Processing and Communication Systems, ICSPCS’2010 - Proceedings, 1–5. https://doi.org/10.1109/ICSPCS.2010.5709752, 2010.