Modelling Synthetic Non-Linguistic Utterances for Communication in Dialogue

Ahmed KHOTA, Asako KIMURA and Eric COOPER

Graduate School of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan

Abstract: Non-Linguistic Utterances (NLUs) present a potentially useful additional communication channel between humans and machines. NLUs may be appropriate in situations such as assisting tourists with various language backgrounds and needs. An experiment was done to model NLU interpretation as dialogue parts. 1000 sounds were generated and evaluated for dialogue relevance to reduce the number to 53, used in the NLU evaluation experiment. 31 Subjects listened to the sounds and rated the following communicative acts for each on a 5-point scale: Greeting, Reject, Question, Thanking, Accept, Apology, Non-Understanding, and Exclamation. Factor analysis yielded three factors: 1) Interrogative vs Apologetic, 2) Appreciative vs Negative, and 3) Positive vs Non-Understanding. Prosodic features such as average pitch, pitch contour, timbre, and duration influenced the interpretation of the sounds. Future work will apply spectral-based speech and emotion recognition techniques along with prosodic features to allow machines to augment dialogue with effective NLUs.

Keywords: *Non-linguistic utterances, NLU, Human-robot interaction*

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. PRELIMINARY EXPERIMENT

3.1 Purpose and methodology

The purpose of the preliminary experiment was 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 were 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.

In the experiment, the random NLU sounds were played for the subjects. 28 audio clips in total are played one after the other, and the subjects selected 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. [7] contains more details.

3.2 Results

10 subjects participated in the experiment, 2 females, 8 males, aged 25–30, from 7 nationalities [7].

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.

4. NLU EVALUATION EXPERIMENT

4.1 Purpose and methodology

The purpose of this experiment is to determine people's interpretations of Non-Linguistic Utterances in terms of their dialogue part. This data would be used to develop the model for synthetic Non-Linguistic Utterances for dialogue augmentation.

Changes were made to the random sound generator based on the results of the preliminary experiment. The pitch range was adjusted so that a lower maximum pitch was used when generating the sounds. The number of Oscillators was increased to 5. This was due to comments from the preliminary experiment that the sounds were too similar in terms of timbre.

To cover the sample space more thoroughly, 1000 random sounds were produced at first. The sounds were produced in the same way as in [7]. These were random in terms of their pitch, amplitude envelope, and timbre, as well as number of pitch changes (quasons) and lengths.

The 1000 sounds were self-tested, and 567 sounds were taken to the next stage of testing. These sounds were tested by three subjects, who were asked to rate each sound on a scale from 1–5, in terms of how much the sound sounded like any part of dialogue.

Each sound was ranked in terms of its Median and Average rating. The top 53 sounds were selected to be used in the main experiment. This number was chosen because in testing it was found that listening to this number of sounds, together with the amount of questions formulated, would take each subject between 20–30 minutes. It was decided that any longer than this might be too long for each subject.

A c# program was written to run the experiment. Figure 1 shows the program. The program plays the 53 sounds in a random order for each subject. The subject begins by entering their Age, Nationality, Home Language, and Gender. For each sound they are asked to rate 8 communicative acts on a 5-point scale, where 1 is strongly disagree, 2 is disagree, 3 is neutral, 4 is agree, and 5 is strongly agree. Each of the 8 communicative acts correspond to dialogue parts from the DAMSL

Modelling Synthetic Non-Linguistic Utterances for Communication in Dialogue

Figure 1: Main Experiment program

model [9,11]. They are each shown with an example in brackets, they are:

Greeting (Hello), Reject/Disapprove (No), Question (What?), Thanking (Thank you/Thanks), Accept/Approve (Yes), Apology (Sorry), Non-Understanding (I don't know), and Exclamation/Surprise (Ah!!)

The program uses a text file called 'Status.txt' to keep track of progress, in case the program is terminated and the subject wants to continue later. The demographic information is saved to another text file called 'Subject.txt', and the rating data per sound is saved to a text file called 'Ratings.txt'. The data from these files is used in the analysis which follows.

4.2 Results

31 subjects (Aged 18–39, 25 Males, 6 Females) in total participated in the experiment. Of those, 23 performed the experiment under controlled conditions in the Laboratory, and 8 performed it remotely. The Laboratory and Remote data were combined after analysis showed that correlations between the two sets of data, as well as 5 pairs of random samples of 23 subjects' data and 8 subjects' data respectively, was good.

26% of the subjects were from South Africa, with English being their home language, 23% of the subjects were from China, with Chinese as their home language, 16% were from Indonesia, with Indonesian as their home language.

Cronbachs alpha was calculated for the data for each of the communicative acts. In each case the value was above 0.9. This indicated a very high level of internal data consistency.

Figure 2: Average Ratings per communicate act

Figure 2 shows the average ratings for each communicative act for all subjects and all sounds. The highest average rated communicative act was Question, while Thanking was the lowest.

Figure 3 shows the sounds for which each had an average rating for any given communicative act higher than 3.5. These 12 sounds can be the most recognizable.

Factor Analysis was done to reduce the number of output variables in terms of the most important contributing communicative acts.

Table 1 shows the results after varimax rotation.

Three factors were yielded. The first factor was made up of Exclamation/Surprise and Question in one direction, and Apology in the opposite direction. This factor is called Interrogative vs Apologetic. The second factor was made up of Thanking and Greeting in one direction and Reject in the opposite direction. This factor is called Appreciative vs Negative. The third factor was made up of Accept in one direction, and Non-Understanding in the opposite direction. This factor is called Positive vs Non-Understanding.

The Factor scores were compared to Average pitch, Duration, number of quasons, pitch contour, and timbre.

Table 1: Factor Analysis results

Communicative Acts	Interrogative vs Apologetic	Appreciative vs Negative	Positive vs Non-Understanding
Thanking (Thank you/Thanks)	-0.14	0.77	0.20
Greeting (Hello)	0.18	0.63	0.41
Non-Understanding (I don't know)	-0.21	-0.09	-0.97
Reject/Disapprove (No)	-0.45	-0.76	0.05
Question (What?)	0.78	0.10	0.04
Accept/Approve (Yes)	0.36	0.37	0.57
Apology (Sorry)	-0.77	-0.23	-0.34
Exclamation/Surprise (Ah!!)	0.81	0.03	0.42

Figure 3: Most recognized sounds

Figure 4 shows that sounds with higher average pitch were more Interrogative, while those with lower average pitch were more Apologetic. The correlation value was 0.53.

Figure 5 shows that sounds with higher average pitch were more Appreciative, while those with lower average pitch were more Negative. The correlation value was

Figure 4: Avg Pitch – Interrogative vs Apologetic

Figure 5: Avg Pitch – Appreciative vs Negative

0.41. There were some outliers in the data, these were the sounds with highest average pitch.

Figure 6 shows that sounds with an upward pitch contour were more Interrogative, while those with a downward pitch contour were more Apologetic.

Figure 7 shows that sounds with an upward pitch contour were more Appreciative, while those with no pitch contour i.e. only a single quason, were more Negative.

Figure 8 shows that sounds with no pitch contour i.e. only a single quason, were more Positive, while those with an updown pitch contour were more Non-Understanding.

Figure 9 shows that sounds with more complicated timbre (indicated by higher number on x axis) were more Appreciative, while sounds with simpler timbre (indicated by lower number on x axis) were more Negative.

Figure 10 shows that sounds with simpler timbre were more Positive, while sounds with more complicated timbre were more Non-Understanding.

Figures 11 and 12 show that sounds with less quasons, and shorter duration were more Positive, while sounds with more quasons, and longer duration were more Non-Understanding (correlation of -0.61 in the latter case).

Figure 13 shows the Amplitude Envelope, which was determined by Attack, Decay, Sustain and Release parameters (ADSR). These values determined the envelope shape as a proportion of the normalized 100% loudness rectangle that represented the maximum amplitude of the sound over its duration. This value for average loudness was also analyzed but it was found to have no effect on the interpretation of the sounds.

Modelling Synthetic Non-Linguistic Utterances for Communication in Dialogue

Figure 6: Pitch Contour – Interrogative vs Apologetic

Figure 10: Timbre – Positive vs Non-Understanding

Figure 7: Pitch Contour – Appreciative vs Negative

Figure 11: #Quasons – Positive vs Non-Understanding

Figure 8: Pitch Contour – Positive vs Non-Understanding

Figure 12: Duration – Positive vs Non-Understanding

Figure 9: Timbre – Appreciative vs Negative

Figure 13: ADSR Envelope

5. DISCUSSION

Looking at the results of this experiment, and comparing them with those of the preliminary experiment, some observations can be made.

Firstly, Although the factor patterns were different, many of the same communicative acts rated prominently in each Factor Analysis. The common ones were: those tags that were Negative, such as Reject in the main experiment, and Negative, Disapproval in the preliminary experiment. Those that were Interrogative, such as Question in both experiments, and those that were Positive, such as Positive, Approval, Greeting and Thanking in the preliminary experiment, and Thanking, Greeting, and Accept in the main experiment.

Then, looking at the correspondence of prosodic features to those factors, there were again some common conclusions reached, namely: That lower average pitches and simpler timbres resulted in sounds that were perceived to be more Negative in their communicative intention, and sounds with upward pitch contours were more Questioning.

In the main experiment, it was not repeated that sounds with downward pitch contours were more Negative.

6. CONCLUSIONS AND FUTURE WORK

Sounds with higher average pitch and upward pitch contours were more Interrogative, while sounds with lower average pitch and downward pitch contours were more Apologetic.

Sounds with somewhat higher average pitch and upward pitch contours, as well as more complicated timbres, were Appreciative, while sounds with lower average pitch, no pitch contours and simple timbres were more Negative.

Sounds with no pitch contours, shorter duration, less quasons, and simpler timbre were more Positive, while sounds with an updown pitch contour, longer duration, higher number of quasons, and more complicated timbre were more Non-Understanding.

Amplitude Envelope or average loudness had no effect on interpretation of the sounds.

Future work will develop the model for interpreting NLUs in terms of dialogue parts. Mel Frequency Cepstral Coefficients, used primarily in speech and emotion recognition [8, 10], will be calculated for each sound.

This data will be fed into a Sparse Encoder Neural Network in order to reduce the MFCC data. Then, the reduced MFCC data along with the experimental dialogue

interpretation results, and prosodic features, will be input into one Neural Network per Communicative act (8 in total) to model and predict the meaning of a new NLU. This would provide a method for generating NLUs for specific applications such as tourist assistance.

REFERENCES

1. Yilmazyildiz, S., Read, R., Belpaeme, T., and Verhelst, W.; Review of semantic-free utterances in social human-robot interaction, *International Journal of Human-Computer Interaction*, 32(1), pp.63-85, 2016.
2. Belpaeme, T., Baxter, P., de Greeff, J., Kennedy, J., Read, R., Looije, R., Neerinx, M., Baroni, I., and CotiZelati, M.; 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, pp.452-459, 2013.
3. Read, R., and Belpaeme, T.; Interpreting non-linguistic utterances by robots, *Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments - AFFINE '10*, pp.65-70, 2010.
4. Read, R., and Belpaeme, T., 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. Read, R., and Belpaeme, T.; People interpret robotic non-linguistic utterances categorically, *International Journal of Social Robotics*, 8(1), pp.31-50, 2016.
6. De Gorostiza Luengo, J. F., Martin, F. A., Castro-Gonzalez, A., and Salichs, M. A.; Sound synthesis for communicating nonverbal expressive cues, *IEEE Access*, 5, pp.1941-1957, 2017.
7. Khota, A., Kimura, A., and Cooper, E.; Modelling of non-linguistic utterances for machine to human communication in dialogue, *International Symposium on Affective Science and Engineering ISASE2019*, pp.1-4, 2019. doi.org/10.5057/isase.2019-C000037
8. Hossan, M.A., Memon, S., and Gregory, M.A.; A novel approach for MFCC feature extraction, *4th International Conference on Signal Processing and Communication Systems, ICSPCS'2010*, pp.1-5, 2010. doi.org/10.1109/ICSPCS.2010.5709752
9. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., and Meteer, M.; Dialogue act modeling for automatic tagging and recognition of conversational speech, *Computational Linguistics*, 26(3), pp.339-373, 2000.

10. Bitouk, D., Verma, R., and Nenkova, A.; Class-level spectral features for emotion recognition, *Speech Communication*, 52(7-8), pp.613-625, 2010. doi.org/10.1016/j.specom.2010.02.010
11. Allen, J., and Core, M.; Draft of DAMSL: Dialog act markup in several layers. Dagstuhl Workshop, 1997.

Ahmed KHOTA (Non-member)

Ahmed Khota received his BSc Engineering (Industrial) from University of the Witwatersrand in Johannesburg, South Africa in 2009. After some years of work in industry, including jobs at Toyota South Africa and a few consulting roles, he went to Japan and has now completed his Master of Engineering at Ritsumeikan University in 2019. Planning on pursuing a PhD at Ritsumeikan University from September 2020, he is currently back in South Africa doing freelance and contract work in industry. His research interests include Affective Engineering, Computer Graphics and Machine Learning.

Asako KIMURA (Member)

Asako Kimura received B.S., M.S., and Ph.D. degrees in engineering science from Osaka University, Suita, Osaka, Japan, in 1996, 1998, and 2001, respectively. She is a Professor at the College of Information Science and Engineering, Ritsumeikan University, Kusatu, Shiga, Japan. She taught at Osaka University from 2000 to 2003 before coming to Ritsumeikan University. She is a member of ACM, IEEE, the Human Interface Society, the Virtual Reality Society, the Information Processing Society of Japan, the Institute of Electronics, Information and Communication Engineers.

Eric W. COOPER (Member)

Eric W. Cooper is a professor in the College of Information Science and Engineering (CISE), Ritsumeikan University. He is a graduate of Maryland Institute College of Art, General Fine Arts (1992) and Ritsumeikan University Graduate School Department of Engineering, Master of Engineering, Information Systems Course (1998). Cooper received his PhD in Engineering from Ritsumeikan University in 2002. He has been a Professor in CISE since 2013 and head of the Affective Engineering and Computer Arts Laboratory (AECAL), in the Information Systems Science and Engineering (ISSE) course since 2017. His research interests are color design support systems, cognitive ergonomics, and affective engineering for interactive systems.