Audience response system-based assessment for analysis-by-synthesis

Jens Edlund[1], Christina Tånnander[2], Joakim Gustafson[1]

[1]KTH Speech Music and Hearing, Stockholm, Sweden
[2]Swedish Agency for Accessible Media (MTM), Stockholm, Sweden
edlund@speech.kth.se, ChristinaTannander@mtm.se, jocke@speech.kth.se

ABSTRACT

We propose a variety of Hollywood's film screenings as a productive tool for phonetic and prosodic research through analysis-by-synthesis. An initial study where the method is used to allow a potential target audience, rather than trained experts, to point out oddities in an extended stretch of connected synthesised speech is presented as proof-of-concept.

Keywords: analysis-by-synthesis

1. INTRODUCTION

In the 1950s, Gunnar Fant and his peers investigated phonetic phenomena – most notably the tie between formants and vowel quality – by implementing their intuitions and hypotheses in speech synthesizers and evaluating the results. To date, speech technology, and speech synthesis in particular, is a productive instrument in phonetic research through analysis by synthesis.

As studies of speech production and perception move from isolated words and utterances in laboratory environments towards connected speech-in-interaction in the wild, the method faces problems with the increasing complexity of both the implementation of hypotheses into synthesis and the evaluation of the resulting synthesized speech. Here, we turn to the problem of assessing longer stretches of connected speech without having to divide it into small isolated segments, compromising the ecological validity of the assessment.

We propose Audience Response Systems (ASR) as a means of tapping into the layman's perception of specific aspects of connected speech. The basic principle of an ARS-based evaluation is that a group of subjects are presented with a continuous stimuli, such as a film. Each subject is equipped with a button and instructed to press it during certain circumstances.

The button-pressing is designed to be physically and cognitively effortless, and usually one single button is used, and the instruction of what subjects should react to is kept open and simple: "Press the button anytime you don't like what you see/hear". To simplify matters further, subjects are told that they can press the button as often or as seldom as they like. Keeping the task simple and effortless allows the subjects to maintain their focus on the stimuli for long stretches of time, and lowers the impact of the side task (i.e. to press the button) on the main task (i.e. to watch/listen to the stimuli).

3. METHOD

3.1 The ARS based synthesis evaluation system

Following up on initial work where we used a web-based test inspired by Audience Response Systems [1], we use a newly designed system for plenary perception tests on streaming data (e.g. auditory, visual, or multimodal streams), built using eight wireless Microsoft Xbox 360 controllers connected to a single computer running custom software that captures button presses. The system runs on custom software that logs each key press with a timestamp and a unique token identifying the console. The system is described in some detail in [2].

3.2 Subjects

For this study, a group of 20 computer science students was used. As our current system can process eight simultaneous subjects, they were divided into two groups of eight and one group of four subjects. Ages ranged from 20 to 30, and 8 of the subjects were female.

The lack of balance in the group (e.g. age, gender, background) was deemed acceptable as the study is intended as proof-of-concept to show the feasibility of the methodology, rather than to validate a particular hypothesis of speech production/perception.

The subjects were, however, a homogenous and typical group of graduate students, and as such represent a representative target audience for that demography.

3.3 Stimuli

The stimulus was a 174 seconds long contiguous excerpt from a synthesised university level text book, using an in-house unit selection synthesis from MTM [3] with the female Swedish voice Tora. The duration of this clip – roughly three minutes – is long enough for it to potentially contain numerous

events that may be deemed exceptional or interesting by listeners. It contained 25 read sentences of an average duration of 6.5 seconds.

We included a deliberate artefact in the stimulus: 55 seconds into the synthesis sequence, between two read sentences, we deleted a synthesis fragment of 200 ms and the following inter-sentence pause. The manipulation did not cause any major audible artefact, such as a sharp clicking noise, but made the end of the former of the sentences very hard to understand as it ended abruptly. This manipulation was performed as a sanity check: if the method works, the manipulation should be clearly visible in the subjects' responses.

3.4 Procedure

The subjects were gathered in a lecture room, handed consoles, and asked to listen to the synthesized text book excerpt while pressing a key in response to a very open question: whenever they heard something they perceived as an error or simply something they did not like. They were told explicitly that the definition of error was left for them to decide upon. They were also told not to worry about their response time, their amount of clicks, or their frequency of clicks, as this would not in itself affect the results. When the instructions had been given, the subjects were asked to be prepared, and the synthesis was then played in high quality loud speakers in the room.

3.5 Analysis

We performed analysis on the experiment data in order to answer the following questions, each of which relies on the previous one being answered in the positive:

1. Are the clicks generated by the subjects distributed in such a manner that we can discern clear peaks where many subjects clicked near-simultaneously?
2. Do the peaks correlate with some known, objective measure of temporally local synthesis quality?
3. Can we find the average response latency – the duration between a problem and the click?
4. Can a professional speech synthesis developer tell what the problems singled out by the peaks are?
5. Does peak height correspond to the gravity of the problems?

If these questions can all be answered in the positive, the system can single out relatively precise segments that are perceived in a certain manner (in this case, as problematic from a quality point of view) by the target audience, suggesting that the system is indeed useful for the analysis step of analysis by synthesis.

The first step in the analysis was to normalize the weight of each click such that the total influence of each subject is the same. For example, each click from a subject producing 50 clicks in total is worth one fifth of each click from a subject producing 10.

Next, Kernel Density Estimation (KDE) over weighted click count and time of click was used to find the areas with the highest density of clicks. Here, for each click, a 250 ms wide Gaussian (roughly the length of a syllable) was added to the KDE at the time of the click, weighted by the originating subject's click rate. In this analysis, no peak finding algorithm was used; peaks were located by visual inspection of the KDE curve.

Objective estimates of synthesis quality were acquired based on the internal state of the unit selection synthesis engine at each phoneme. Two sets of information were used to produce KDE curves corresponding to the click-based curves: (1) a list of phoneme borders with information about whether a cut occurred at each border and (2) a list of the number of mismatches for each phoneme out of the following contextual matching criteria: target phone, content/function word, phone position in syllable and phrase, syllable position in word, previous and following target phones, language (Swedish or English) and whether the target word should be spelled out or not. KDE curves were produced using the same configuration as before, but with these objective performance criteria as input. The manually inserted artefact described in 3.3 was not given an objective score.

It was known from precious tests with the same ARS-based system ([2]) that the minimum response latency (reaction time + system latency) was up towards 500 ms, so the perceived problem must be located somewhere 500+ ms prior to each peak. The correspondence between the click-based curves and the curves based on the internal state of the synthesis engine was measured by finding the first peak in the cross-correlation between the two with a latency of at least 500 ms. The average latency for this task was estimated by finding the latency of the same peak in the cross-correlation.

The click-based KDE curve was then realigned to adjust for the average latency, so that each peak would be maximally probable to be temporally co-positioned with the problem that caused the subjects to click.

A professional speech synthesis developer then judged the speech synthesis in the vicinity of each peak with respect to how easily identifiable the problem was on a scale of 1 through 3, where 1 meant "readily identifiable", 2 "identifiable", and 3

"nor clearly identifiable". For the peaks labeled with 1 or 2 (i.e. as identifiable), the temporal distance from the peak to the problem was also noted.

4. RESULTS

4.1 General statistics

Over the 174 seconds of synthesis, subjects clicked on average 29 times per subject (i.e. every sixth second), with a range from 10 clicks (i.e. once in 17 seconds) to 50 clicks (i.e. once in 3.5 seconds).

4.2 KDE estimates

The KDE estimation curve over all subjects and the entire stimuli is shown in Figure 1. The 10 or so tall peaks in the curve mark places were a very large proportion of our subjects pressed a button as a reaction to a perceived problem in the synthesis.

Figure 1. Overview of the synthesized waveform with the KDE estimates from all subjects (top pane) and a zoomed-in section of 2.5 seconds (bottom pane). The X axis shows time, the Y axis the relative probability of a click at that time. The darker plot with the slightly higher peak in the lower pane is the subject normalized estimates.

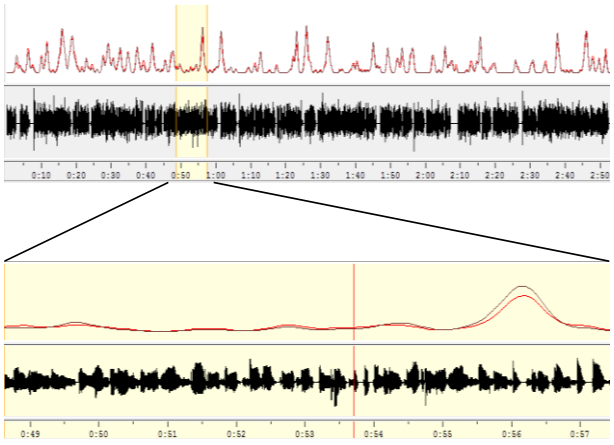
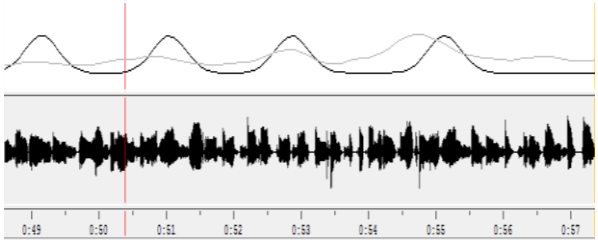


Figure 2 shows the KDE curves based on the synthesis engine's internal state for the same segment as the lower pane of Figure 1.

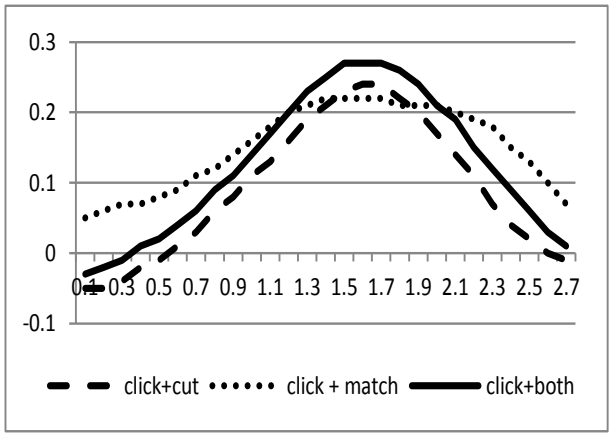
Figure 2. KDE curves based on the presence of a cut at each phoneme border (dark line) and the number of mismatched context criteria at each phoneme (light line) over an eight second segment. The X axis shows time and the Y axis the estimated relative probability of a cut or mismatch in contextual matching criteria at that time.



4.3 Cross-correlations

The first peak of the Pearson cross-correlation between the KDE curve based on subject clicks and the ones based on the internal state of the synthesis (cuts present, number of contextual mismatches present, and a combination of both) was calculated. The results show peaks of small, bordering to medium correlations for all three (0.22, 0.24, 0.27), all at about 1.6 seconds (see Figure 3). As the manually inserted artefact described in 3.3 was not given an objective score, it could not affect the cross-correlation positively.

Figure 3. The first peaks of the cross-correlations between the KDE curve of the subjects' clicks and those based on the presence of cuts, the presence of contextual mismatches, and a combination of both. The X axis shows time, the Y axis the Pearson correlation at that latency.

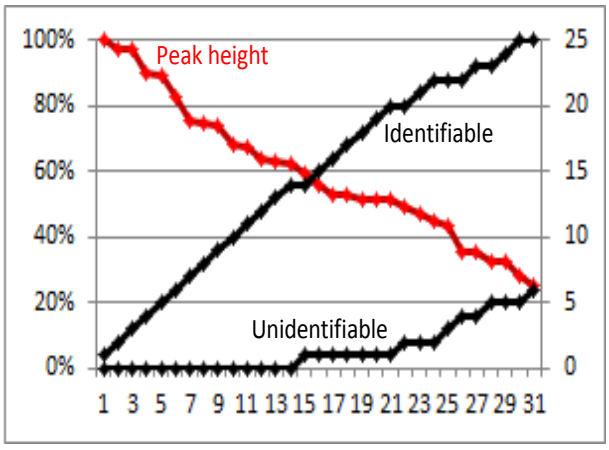


4.4 Manual annotation

Once the click-based KDE curve had been realigned (pushed back by the 1.6 seconds suggested by the cross-correlation), the professional synthesis developer annotated its 31 tallest peaks. Of these 31 peaks, 21 (68 %) were judged to be connected to an easily identified problem, 4 (13 %) were judged to be identifiable, and 6 (19 %) were not found to be connected to a clear problem in the synthesis by the professional developer. Amongst the 21 easily detected problems, the average temporal distance between the realigned click peak and the problem was 0.3 s, or just above one syllable.

If the click-based peaks are inspected in order of their height (see Figure 4), we see that the first unidentifiable problem occurs only at place 15, at which point the peak height has dropped under 60 % of the tallest peak. Put differently, the proportion of identifiable problems over the 14 tallest peaks is 100%, over the first 21 peaks it is 95 %, and it drops to 81 % after all 31 peaks have been inspected. The tallest peak of all is, predictably, the manually inserted artefact described in section 3.3, which is incidentally also the peak shown in the lower pane of Figure 1.

Figure 4. The falling line represents the relative peak height for each of the 31 tallest click-based peaks, in percentage of the highest peak (left axis). The rising lines represent the cumulative number of judgments (right axis). The top rising line represents the identifiable problems (judged as category 1 or 2), and the bottom rising line the unidentifiable problems (judged as category 3).



5. CONCLUSIONS

We have presented a method for assessment of specific aspects of synthesised connected speech. The method stands out in that it (a) allows developers to evaluate one single stimulus, rather than comparing different stimuli, (b) allows subjects to assess the speech in context, (c) does not give subjects a chance to over-think these assessments, (d) has subjects listen to long, continuous stretches of synthesis that better reflect real-world situations, and (e) is cost efficient. For these reasons, the method is well-suited for analysis-by-synthesis of connected speech and conversations. Note that the method should also work well for pointing out events in human speech – we have initiated experiments to this effect.

The results from our proof-of-concept experiment allows us to answer all five questions posted in section 3.5 in the positive: (1) clicks generated by subjects are distributed such that clear peaks can be easily found; (2) peaks correlate with known internal states associated with quality; (3) we can find the average response latency; (4) in the majority of cases, a professional speech synthesis developer can find what likely caused subjects to click, (5) especially for high peaks.

We note that although a professional synthesis developer can find the likely cause of subject clicks, the same results would not have been achieved using only the professional to select the problems to focus on: the consumption of the professional's time would have been much greater, and the professional finds problems that subjects do not perceive or that they choose to ignore. We have shown that our method achieves its goal of pointing out a specific perceptual event – errors – in speech synthesis.

6. ACKNOWLEDGEMENTS

This work was funded in part by the Swedish Research Council (VR) project Incremental Text-To-Speech Conversion (2013-4935).

7. REFERENCES

[1] Tånnander, C. (2012). An audience response system-based approach to speech synthesis evaluation. In The Fourth Swedish Language Technology Conference (SLTC 2012) (pp. 74-75). Lund, Sweden.

[2] Edlund, J., Al Moubayed, S., Tånnander, C., & Gustafson, J. (2013). Temporal precision and reliability of audience response system based annotation. In Proc. of Multimodal Corpora 2013. Edinburgh, UK.

[3] Sjölander, K., Sönnebo, L., & Tånnander, C. (2008). Recent advancements in the Filibuster text-to-speech system. In Proc. of the Swedish Language Technology Conference (SLTC 2008). Stockholm.