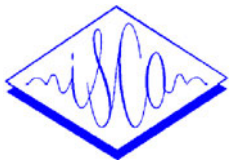

7th ISCA Workshop on
Speech Synthesis (SSW-7)
Kyoto, Japan
September 22-24, 2010

Speech acts and dialog TTS

Ann K. Syrdal, Alistair Conkie, Yeon-Jun Kim, Mark Beutnagel

AT&T Labs – Research
Florham Park, NJ USA
{syrdal,adc,yjkim,mcb}@research.att.com

Abstract

The approach outlined in this paper aims to provide better expressivity of unit selection TTS for dialog intended applications while retaining the natural sounding voice quality typical of unit selection synthesis. A small set of speech acts were used to annotate a corpus from one female US English speaker. The corpus was composed of speech read primarily from interactive dialogs of various kinds. Global acoustic variables related to prosody were calculated for each speech act in the corpus. A hierarchical cluster analysis performed on the acoustic variables showed clustering that corresponded to general classes of dialog speech acts. The acoustic prosodic variables were used to specify pitch range parameters of a unit selection Speech Act TTS voice. Listening tests indicated large and significant improvement in rated speech quality for the Speech Act system compared to the Standard TTS system built from the same speaker.

Index Terms: speech synthesis, dialog, speech acts, prosody

1. Introduction

In the last dozen years, advances such as high quality unit selection synthesis [1] have greatly improved the naturalness of synthetic speech because minimal signal processing has resulted in less distortion. However, the limitations of even high quality general-purpose TTS for human-computer dialogs have become more apparent as natural language dialog systems have advanced in sophistication. The improved naturalness provided by unit selection synthesis has been achieved at the cost of the more precise prosodic control provided by more robotic sounding synthesizers. Since prosody conveys much of the subtlety and complexity of meaning in natural language dialogs, the narrow expressive range of TTS is a major drawback to its use in human-computer dialogs.

Human-computer dialogs are more pragmatic than dramatic, and they rarely involve the expression of such basic emotions as anger, sadness, surprise, disgust, or even happiness. For this reason, we have focused instead on the communicative function of an utterance in an interaction: speech acts. Our goals are (1) relevant and meaningful prosodic variation in dialog applications, (2) more prosodic and expressive control over unit selection TTS while retaining naturalness, and (3) accessibility of prosody control by spoken dialog systems and by non-expert users.

An earlier report of our work [2] focused on the analysis of prosodic features and their relation to speech acts. This paper describes dialog speech acts and acoustic measures of some of their prosodic characteristics, then discusses the construction of a Speech Act TTS system, and finally it describes a listening test of Speech Act TTS and reports its results.

2. Dialog speech acts

Speech acts are intended to classify the purpose or communicative function of an utterance [3], and dialog acts are speech acts in the context of an interactive dialog [4]. We do not claim that the set of speech acts used in our study is exhaustive, nor was it theoretically motivated. As used in our study, most dialog speech acts fall into four broad categories, listed below in Table 1 with counts of instances in the corpus and some examples. Note that the Warning speech act was excluded from analysis due to small sample size.

Table 1: *Dialog Speech Acts*

Imperative: directs actions of others			
Speech Act	Abbr.	Num.	Examples
Request	Req	319	*Please enter your PIN.*
Directive	Dir	459	*Turn left onto Main Street.*
Warning	Warn	7	*Be prepared to stop.*
Repeat	Rept	62	*Pardon me?*
Wait	Wait	121	*Just a second please.*
Interrogative: solicits information from others			
Speech Act	Abbr.	Num.	Examples
Question-wh	Qwh	641	*Who should I call?*
Quest.-yes/no	Qyn	2394	*Are you flying to Cleveland?*
Quest.-mult.choice	Qmc	100	*Downtown or near the airport?*
Assertive: conveys factual information to others			
Speech Act	Abbr.	Num.	Examples
Inform.-detail	Idet	464	*VTL dash help at VT dot net.*
Inform.-general	Igen	4713	*You have four new messages.*
Affective: expresses the speaker's attitude			
Speech Act	Abbr.	Num.	Examples
Greeting	Grt	205	*Hi! Welcome to Call ATT.*
Apology	Apol	355	*I'm sorry.*
Exclam.-negative	Eneg	17	*Oops! Oh dear!*
Exclam.-positive	Epos	16	*Great!*
Thanks	Thks	129	*Thanks for calling.*
Goodbye	Gbye	39	*Bye bye.*
Cue phrase	Cue	349	*Meanwhile, ... Well, ...*
Back-channel	Fill	32	*Hmmm. Uh-huh.*
Other?			
Speech Act	Abbr.	Num.	Examples
Confirmation	Conf	1728	*All right.*
Disconfirmation	Dis	1670	*No, you must change terminals.*

3. Speech corpus

Approximately 12 hours of digitally recorded speech sampled at 16 kHz were used as the corpus for this study. All record-

ings were made using a high quality head-mounted condenser microphone in a nearly anechoic recording room.

The speech corpus was recorded from an adult female who was a native speaker of American English. She was a paid voice talent with professional training and several years of experience as a voice-over artist and actress.

The speaker read text material that we believed would be most useful in human-computer dialog applications. Texts included dialogs that were transcribed from customer-live agent interactions, simulated dialogs based on such interactions, prompts for various interactive services, laboratory sentences for phonetic coverage, and information often requested from automated interactive services, such as names, addresses, flight information, digit strings such as used for telephone, account, or credit card numbers, natural numbers, and letters of the alphabet, used for spelling out words.

The speech act of every utterance in the 12 hour corpus was annotated manually by the first author. Often the text of the utterance and its context was sufficient to determine the most appropriate speech act tag, but some cases required listening to the recorded speech as well. The utterance "Okay" served a variety of dialog functions in different contexts, for example, and often required listening for speech act classification.

4. Acoustic measures of speech act prosody

This paper focuses on relatively global aspects of prosody rather than on phrasing and intonation. The following six acoustic measures of prosody were made based on signal analysis software used in the preparation of a recorded speech inventory for unit selection synthesis.

- Max F0: The maximum F0 value of each speech act utterance was calculated from units that were fully voiced throughout their duration. Because of that constraint, this and other F0 measures are very robust.
- Min F0: The minimum F0 value of each speech act utterance was also calculated from units that were 100% voiced.
- F0 Range: The range was calculated per speech act utterance from its max F0 - min F0.
- Mean F0: The mean F0 of all fully voiced units was calculated for each speech act utterance.
- Mean Phone Duration: The mean duration of all phones (regardless of voicing) that were included in the entire set of utterances tagged with the same speech act. This is a measure of speaking rate (the faster the rate, the shorter the duration).
- Mean Power: The mean log power of all phones (regardless of voicing) included among all the utterances in the same speech act set.

A scatter plot of F0 range (on the y-axis) as a function of mean F0 (on the x-axis) of each speech act is shown in Figure 1. There is wide variation among speech acts in both F0 measures. Mean F0 ranges from a low of 170 Hz for Exclamation-negative (Eneg) utterances to a high of 254 Hz for speech acts classified as Repeat (Rept). Eneg utterances have the narrowest pitch range (15 Hz) of the speech acts, and the widest pitch range was 163 Hz for Requests (Req). Speech acts with low F0 ranges and relatively low mean F0 include Eneg, Gbye, Cue, Fill, Apol, Thks, and Idet. Speech acts with higher pitch ranges and often higher F0 means include Igen, Dir, Dis, Wait, Req, Conf, Qmc, Qyn, Qwh, Grt, Epos, and Rept.

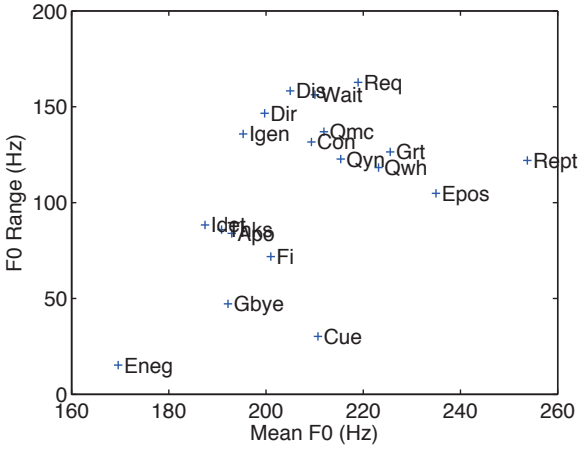

Figure 1: *Pitch range and mean F0 of speech acts.*

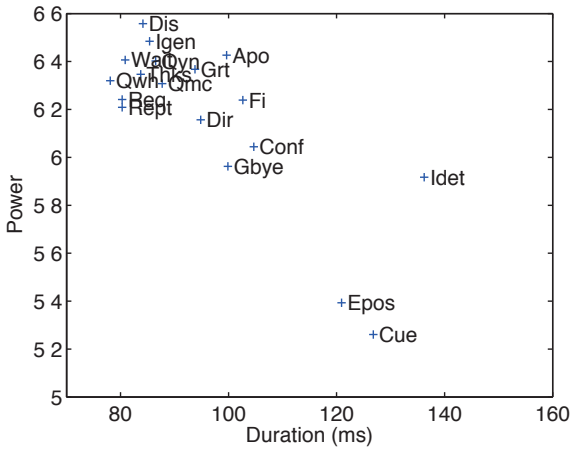

Figure 2: *Average phone duration and log power of speech acts.*

Figure 2 is a scatter plot of the average phone duration (in ms) and average log power for each speech act. Note that because of its extremely long average phone duration (234 ms) and its log power at 4.4, the Eneg speech act was omitted from the plot to more clearly show the distribution of the other speech acts. The fastest speaking rate was observed for wh-questions (Qwh) as indicated by an average phone duration of 78 ms. Log power also differentiated some speech acts from others. Exclamations, both positive (Epos) and negative (Eneg), as well as Cue utterances had by far the lowest log power, and Disconfirmations (Dis), the highest.

There are large differences in speaking rate (Figure 2) and F0 range (Figure 1) between the two Assertive speech acts: Informative-general (Igen) and Informative-detail (Idet). Mean phone duration was 85 ms for Igen but 136 ms for Idet, indicating that the talker slowed her speaking rate down considerably when reading detailed, information-dense material. The pitch range was considerably higher for Igen (136 Hz) than for Idet (88 Hz) utterances, although the F0 means differed by less than 8 Hz.

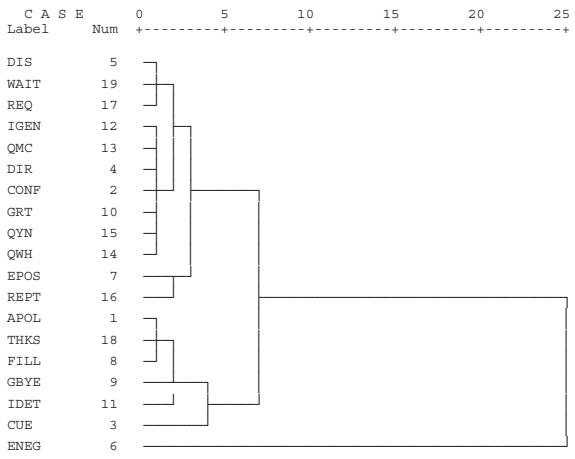

Figure 3: *Hierarchical clustering dendrogram of speech acts.*

A hierarchical cluster analysis was performed on the basis of all six acoustic measures of the 19 dialog speech acts. The results of the cluster analysis are presented in the form of a dendrogram, shown in Figure 3. A dendrogram is a tree diagram that illustrates the arrangement of the clusters produced by a clustering algorithm. The left column of nodes represent the 19 speech acts, arranged according to pairwise similarity: adjacent speech acts are more similar than distant ones. The nodes of the tree diagram represent the clusters to which the speech acts belong, and the horizontal length of the lines represents the distance between clusters.

Exclamation-negative (Eneg), at the bottom of the tree, is differentiated at an early stage of clustering from all the other speech acts. At the second node in the dendrogram, at the top of the tree, all the Imperative and Interrogative speech acts fall into a large cluster along with Informative-general (Igen), Disconfirmation (DIS), Confirmation (CONF), and the two most emotionally positive Affective speech acts, Greeting (Grt) and Exclamation-positive (Epos). The remaining Affective speech acts, Apology (Apol), Thanks (Thks), Good-bye (Gbye), Backchannel (Fill), and Cue phrase (Cue) fall within the lower cluster formed by the split at the second dendrogram node. These speech acts all tend to be quite scripted and passive in nature.

5. Dialog speech acts applied to TTS

We have implemented a prototype dialog speech act TTS system that sets global prosodic variables according to the speech act specified. The acoustic inventory of the unit selection system consists of the 12-hour corpus described above. An evaluation of this prototype system is described below.

6. Listening tests

Two web-based listening tests were conducted to test whether the use of a specialized Speech Act TTS system improved the subjective quality of synthetic speech in the context of a human-computer dialog. The first test was conducted following our normal procedure with AT&T Labs Research employees serving as listeners; many of the participants were speech researchers, but the majority were not. Listeners for the second test were self-selected volunteers who responded to an invitation posted on a website. The second test was run to compare results from the first test with results from a larger, less controlled, and relatively anonymous group of listeners, and to explore the validity and feasibility of such testing.

6.1. Stimuli

The automated agent portion of a simulated dialog in a travel reservations IVR scenario was synthesized using two different TTS systems: (1) Standard TTS (Std) used the standard AT&T Research unit selection TTS system, and (2) Speech Act TTS (SpAct) used the Speech Act experimental system described above. Both systems were built from the same speaker, although the recorded material included in the two inventories differed in both size and constituent material. The standard TTS inventory contained approximately 6 hours of speech; the recorded material was primarily reading of factual material, but it also included some interactive dialog material. From each system, seven utterances (representing agent turns in a dialog) were generated and used as listening test stimuli. Table 2 lists the speech acts that composed each of the seven test utterances.

Table 2: *Speech acts included in test utterances.*

Utt.	Speech Acts in Test Utterance
1	greeting + wh-question
2	confirmation + yes/no question
3	wh-question
4	confirmation + informative-general
5	exclamation-positive + 2 yes/no questions
6	exclam.-positive + inform.-general + inform.-detail
7	thanks + good-bye

The input text was standard text for Std TTS. The text was annotated for SpAct TTS with mark-up indicating speech act and its associated mean pitch range.

6.2. Method

The initial web-based listening test had two parts: (1) seven paired comparisons in which listeners rated their A/B preference on a -2 (strongly prefer A) to +2 (strongly prefer B) scale, where 0 indicated no preference. Order across the seven pairs was randomized, and A/B position within each pair was counter-balanced across listeners. (2) The 7 utterances synthesized by each TTS system were concatenated (with a beep between utterance turns) into a single audio file and listeners were asked to rate the overall quality of each of the resulting two files on a 5-point scale (1=Bad to 5=Excellent). Two comparison pairs were used for practice, so listeners could adjust their audio level and become familiar with the paired comparison procedure.

The second listening test was also web-based but only contained three paired comparisons, selected on the basis of results from the first test. The utterance pair which resulted in the most favorable score for Speech Act TTS, the pair with the least favorable score for Speech Act TTS, and the utterance pair with the median score were included in the second test.

In both tests, listeners also indicated whether or not English was their native language, and whether they listened using headphones or speakers.

6.3. Listeners

In the first test, 83 listeners (all AT&T employees) participated in the test; 49 (59%) were native speakers of English, and 34

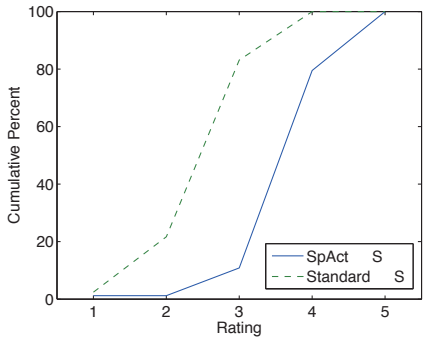

Figure 4: *Cumulative Distribution of Subjective Quality Ratings*

(41%) were non-native speakers; 33 (40%) listened using headphones and 50 (60%) by speaker.

In the second test, 500 anonymous listeners volunteered to participate; 398 (80%) reported to be native speakers of English, and 102 (20%) reported they were non-native; 180 (36%) listened via head/ear phones, and 320 (64%) by speaker.

6.4. Results

6.4.1. Test 1

Table 3 lists Comparison Mean Opinion Scores (CMOS) from paired comparison ratings of each of the seven dialog agent turns. Scores ranged from -2 (Std much preferred) to +2 (SpAct much preferred). A score of 0 indicated no preference. The table also lists standard deviations and standard errors of the mean.

Table 3: *Test 1. CMOS per Utterance Pair*

Pair	N	Mean	SD	SE
P1	83	1.060	0.888	.098
P2	83	1.169	0.998	.110
P3	83	0.759	1.089	.119
P4	83	0.940	1.097	.120
P5	83	1.420	0.650	.071
P6	83	0.614	1.046	.115
P7	83	0.120	1.109	.122

A One-Sample T-Test was run on ratings for each of the seven utterance pairs, testing (two-tailed) the null hypothesis of 0 (no preference). Utterance pairs 1 - 6 were significantly different from (all higher than) 0, but the score for pair 7 did not differ significantly from 0.

A Repeated Measures ANOVA was conducted with Utterance Pair (7) the within-subjects factor, and Language (2) and Listening Mode (2) the between-subject factors. There was a significant main effect of Utterance Pair ($F=20.565(6,74)$, $p < 0.0001$), but no other significant effects or interactions. Post-hoc comparisons of Utterance scores indicated that the pair 5 was significantly higher than all others, and pair 7, significantly lower. Among the remaining five pairs, pairs 2, 1, and 4 did not differ from one another, nor did pairs 1, 4, 3, and 6.

Overall quality ratings of the entire dialog sequence of seven utterances synthesized by the Std and SpAct TTS systems were also analyzed. Figure 4 shows the distribution of ratings for the two TTS systems, displayed as a cumulative percentage. The Mean Opinion Score (MOS) of Std TTS was 2.99, while the SpAct MOS was 4.09.

A Repeated Measures ANOVA was run on the overall quality ratings with TTS System (2) the within-subject factor, and Language (2) and Listening Mode (2) the between-subject factors. There was a significant main effect of TTS System: SpAct TTS scores were significantly higher than Std TTS scores ($F=150.034(1,79)$, $p < .0001$). The TTS System x Language interaction approached significance ($F=3.843(1,79)$, $p < .053$). Native English speakers rated Std TTS .28 MOS lower, and SpAct TTS .09 MOS higher, than non-native English speakers. No significant differences were found between listeners who used headphones versus speakers.

6.4.2. Test 2

Table 4 lists Comparison Mean Opinion Scores (CMOS), standard deviations, and standard errors of the mean from paired comparison ratings of each of the three utterance pairs included in Test 2.

Table 4: *Test 2. CMOS per Utterance Pair*

Pair	N	Mean	SD	SE
P4	500	1.212	1.129	.050
P5	500	1.582	0.832	.037
P7	500	-.542	1.373	.061

The same statistical analyses of preference ratings described for Test 1 were repeated on the larger Test 2 data set. One-Sample T-Tests found that means of all three utterance pairs were significantly different from 0; pairs 4 and 5 were higher, and pair 7 was lower. Pair 7 results differed from those in Test 1, where the mean was not significantly different from 0. In Test 2, the negative score indicated a small but significant preference for the Std TTS version of utterance 7.

Order bias in favor of the second member (B) of an A/B pair is often observed in non-interactive paired comparison listening tests in which stimuli are presented to listeners in a fixed order. A One-Sample T-Test with test value = 0 was conducted to test A/B order bias in our interactive web-based test, in which participants could listen to pairs in any order and as many times as they wished. The test confirmed a significant bias in favor of the second member of the pair ($t=2.517$, $df=1499$, $p < .012$ (2-tailed)). No significant order bias was observed in Test 1.

A Repeated Measures ANOVA was also conducted with Utterance Pair (3) the within-subjects factor, and Language (2) and Listening Mode (2) the between-subject factors. There was a significant main effect of Utterance Pair ($F=253.737(2,992)$, $p < .0001$), and again post-hoc comparisons indicated that pair 5 was significantly higher than the others, and pair 7, significantly lower. There was a significant between-subject effect of Language ($F=7.062(1,496)$, $p < .008$): scores were significantly higher for native than non-native English speaking listeners (means were .785 and .601, respectively). The Utterance Pair x Language interaction was also significant ($F=3.626(2,992)$, $p < .027$); native English speakers had higher scores for Pairs 4 and 5, but lower scores for Pair 7, than non-native speakers. There were no other significant effects or interactions.

7. Summary and conclusions

Speech acts differ greatly among one another along the various acoustic dimensions of prosody measured: maxF0, minF0, F0range, meanF0, speaking rate as measured by phone duration, and power. Speech acts form meaningful groups when hierarchically clustered on the basis of their acoustic measures.

Listening tests indicate that setting the prosodic parameters on the basis of speech act significantly and dramatically improved perceived TTS quality for utterances representative of human-computer dialogs. The composition of the utterances for which preference for SpAct TTS was highest suggests that questions, particularly yes/no questions, may contribute appreciably to improvements in perceived TTS quality.

Test results were notable in two other respects as well, both relevant to listener selection in testing. Firstly, listeners who were native speakers of English were significantly more discriminating in their perceptual judgments than non-native speaking listeners. Secondly, testing a large group of anonymous listeners from the internet yielded results that correspond quite closely to those obtained from a smaller and more carefully controlled listener group.

8. Future directions

Speech act TTS may be coordinated with a spoken dialog system to the advantage of both. In spoken dialog systems used for human-computer dialog, the dialog manager specifies the purpose of an utterance it needs to generate in order to further the dialog. This utterance goal is equivalent to a speech or dialog act. A language generation module then determines the wording of the utterance and normally passes the text generated to a speech synthesis system, which generates audible speech output. A dialog system also can convey the intended speech act to a TTS system designed to use speech act information as well as text in synthesizing speech. Other alternatives to providing speech act information to TTS include analysis of input text to predict the most likely speech act intended or manual text markup.

A TTS front end performs text normalization and syntactic analysis, determines word pronunciation and makes prosodic assignments including phrasing, prominence, intonation contour, and phone durations. Our acoustic analysis of prosody indicates that there are, at least for the speaker studied, systematic differences in pitch, pitch range, phone duration, and power among different speech acts. Although beyond the scope of the current study, there are also cases in which speech acts strongly influence the intonation contour of an utterance [2]. We expect that including speech act information along with input text would improve the capability of a TTS front end to assign several aspects of prosody more appropriately.

9. Acknowledgements

The authors would like to thank the many listeners who voluntarily participated in this experiment.

10. References

[1] M. Beutnagel, A. Conkie, J. Schroeter, Y. . Stylianou, and A. Syrdal, "The AT&T Next-Gen TTS System," in *Proc. Joint Meeting of ASA, EAA, and DEGA*. Berlin: ASA, EAA, and DEGA, March 1999, p. SASCA__4, http://www2.research.att.com/ ttsweb/tts/pubs.php.

[2] A. K. Syrdal and Y.-J. Kim, "Dialog speech acts and prosody: Considerations for TTS," *Fourth International Conference on Speech Prosody*, 2008, http://www2.research.att.com/ ttsweb/tts/pubs.php.

[3] J. R. Searle, *Speech Acts*. London-New York: Cambridge University Press, 1969.

[4] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech." *Computational Linguistics*, vol. 26, no. 3, pp. 1–34, 2000.