

Dialog speech acts and prosody: Considerations for TTS

Ann K Syrdal and Yeon-Jun Kim

AT&T Labs - Research
Florham Park, NJ (USA)

{syrdal; yjkim}@research.att.com

Abstract

As natural language dialog systems involving both speech recognition and text-to-speech (TTS) synthesis become more sophisticated, the limitations of general-purpose TTS for human-computer dialogs have become more apparent. Much subtlety and complexity of meaning in natural language dialogs is conveyed by prosody; how something is said is often as important as what words are spoken. At the same time, advances such as unit selection synthesis have greatly improved the naturalness of synthetic speech because much less signal processing is required, resulting in less distortion. However, the improved naturalness provided by unit selection synthesis has been achieved at the cost of the more precise prosodic control provided by earlier, more robotic sounding synthesizers.

With the goal of providing more prosodic and expressive control over unit selection TTS for dialog applications, while retaining naturalness, we have focused on speech acts, the communicative function of an utterance. The current working set of speech acts being used includes:

- Imperative: directive, request, wait, repeat, warning
- Interrogative: question-wh, question-yes/no, question-multiple choice
- Assertive: informative-general, informative-detail
- Affective: apology, exclamation-positive, exclamation-negative, greeting, good-bye, thanks
- Others: confirmation, disconfirmation, back-channel, cue phrase

Our work is practically focused, but also involves some observations of more general interest. We use a relatively small set of speech acts both to classify utterances in a speech corpus according to their communicative function, and then to preferentially select speech act-appropriate units to match the desired speech act of the utterance to be synthesized. The corpus is composed of speech read (primarily from interactive dialogs of various kinds) by a female US English speaker (a voice talent used to build one of our TTS voices). We examine prosodic differences of a more “global” nature (mean f_0 , f_0 range, speaking rate, energy level) for the entire set of speech acts. A portion of the database has also been ToBI labeled and analyzed for systematic differences. There are several significant prosodic differences among the various speech acts.

In our current TTS implementation, speech acts are being used as another feature to select speech units for concatenation, but results from analyzing prosodic features of the various speech acts will also be used to better predict the prosodic features desired. Results thus far are promising and examples will be demonstrated.

1. Introduction

As natural language dialog systems have become more sophisticated, the limitations of general-purpose TTS for human-computer dialogs have become more apparent. Much subtlety and complexity of meaning in natural language dialogs is conveyed by prosody. At the same time, advances such as unit selection synthesis [1] have greatly improved the naturalness of synthetic speech because much less signal processing is required, resulting in less distortion. However, the improved naturalness provided by unit selection synthesis has been achieved at the cost of the more precise prosodic control provided by earlier, more robotic sounding synthesizers.

With the goals of providing (1) 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 to non-experts, we have focused on the communicative function of an utterance: speech acts. Our work represents a practical approach rather than one that is more theoretically motivated.

There has been much recent interest in expressive TTS, which has generally focused on conveying emotion. Our work on emotional TTS [2] demonstrated to us that both voice quality and prosody have strong and sometimes independent effects on synthesized emotion. However, while it is an interesting and worthwhile goal to synthesize speech that can convey anger, sadness, and happiness, these emotions are not typically relevant for TTS expression in actual human-computer dialogs.

The primary focus of our paper is the acoustic measures of prosody in a large (12-hour) speech corpus from one speaker and their relation to speech acts. These more global aspects of prosody have been studied relatively less than phrasing and intonational features. The acoustic measures studied include maximum F_0 , minimum F_0 , pitch range, and mean F_0 per utterance, phone duration, and power. Hierarchical cluster analysis of speech acts was then performed based on the six acoustic measures.

A large scale analysis of intonational features was beyond the scope of this paper, but examples are presented demonstrating the influence of speech acts on intonational contours.

Finally, we briefly discuss the application of speech acts to unit selection TTS in dialog applications.

2. Methods: Speech act classification and acoustic measurements of dialog prosody

This methodology section describes the recorded speech corpus, speech act categories and annotation, and the acoustic measurements and ToBI annotation of prosody.

2.1. Speech corpus

Approximately 12 hours of digitally recorded speech (recorded at 48 kHz and subsequently downsampled to 16 kHz for analysis) were used as the corpus for this study. All recordings were made using a high quality head-mounted condenser microphone in a nearly anechoic recording room.

2.1.1. Speaker

The corpus was recorded from one young adult female 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.

2.1.2. Recording material

The speaker read from various texts. Texts included dialogs that were transcribed from actual customer-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. Material that we believed would be most useful in human-computer dialog applications was included in the corpus.

2.2. Speech acts

Speech acts are intended to classify the purpose or communicative function of an utterance in a dialog. The set of speech acts used in our study does not claim to be exhaustive nor was it theoretically motivated. It was arrived at from the practical perspective of trying to use a relatively small set of categories to identify the basic goal or function of the utterances in our speech corpus. The speech act of each utterance in the corpus was tagged 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.

The set of speech acts used in the current study are listed below. Most of the speech acts are readily classifiable into one of the four broader modes of enunciation reflecting the speaker’s cognitive attitude to the content: Imperative, Interrogative, Assertive, and Affective. A few of the speech acts were somewhat ambiguous, and they are listed below under “Other”. Examples and the number of each speech act tagged in the corpus (in parentheses) is included in the listing.

- Imperative

- Directive (Dir). Examples: “Record at the tone.” “Just give me a call tomorrow by five.” (459)
- Repeat (Rept). Examples: “Could you repeat that?” “Pardon me?” “What was that again?” “I didn’t get that.” (62)
- Request (Req). Examples: “Please say yes or no.” “Let’s return to this later.” “Please enter your pin.” (319)
- Wait (Wait). Examples: “Please hold. I’m still waiting to hear back.” “Please wait. It will be just

a minute.” “Let me check. Just a second please.” (121)

- Warning (Warn). Examples: “Be prepared to stop.” “narrow bridge.” “no left turn.” (7: excluded from analysis due to small sample size)

- Interrogative

- Question-multiple choice (Qmc). Examples: “What kind of car did you prefer, a compact car or a standard car?” “Do you prefer to stay at the Hilton near the airport or downtown in Los Angeles?” (100)
- Question-wh (Qwh). Examples: “How may I help you?” “Who should I call?” “What’s next?” “What time would you like to return?” (641)
- Question-yes/no (Qyn). Examples: “Should I send it?” “Are you flying to Cleveland?” “Did you need a hotel in Chicago?” (2,394)

- Assertive

- Informative-detail (Idet). Examples: “VTL, dash technical, dash help at voicetone dot net.” “the address is three four oh one south thirty five, Austin Texas, seven eight seven four one.” (464)
- Informative-general (Igen). Examples: “Four new messages.” “You may have chosen a city with limited schedules or days when nothing is scheduled.” “I have economy cars, compact cars, luxury sedans and convertibles.” (4,713)

- Affective

- Apology (Apol). Examples: “I’m sorry.” “Sorry.” “Sorry, we seem to have a bad connection.” “I am sorry, but the computer just went down.” (355)
- Exclamation-negative (Eneg). Examples: “oh!” “oops!” (17)
- Exclamation-positive (Epos). Examples: “Great!” “Fantastic!” “Excellent!” “That’s good!” “Wow!” “That’s it!” “Good news!” “Cool!” “Okay!” (16)
- Goodbye (Gbye). Examples: “Bye bye.” “Good bye.” “Hope to hear from you again.” “We look forward to seeing you at the winter two thousand Innovation Forum.” (39)
- Greeting (Grt). Examples: “Hi, this is Annie.” “Welcome to call ATT.” “It’s nice to hear from you.” “Hello. AT&T Communicator.” “hi! how are you?” (205)
- Thanks (Thks). Examples: “Thank you.” “Thank you very much.” “Thanks for calling AT&T Communicator.” “Thank you for taking part in the evaluation.” “You’re welcome, Edward.” “That’s no problem.” (129)

- Other

- Confirmation (Conf). Examples: “Got it.” “OK.” “All right.” “Window seating preference noted.” “Okay then.” “Yes.” “Okay, you’re all set.” (1,728)
- Cue phrase (Cue). Examples: “Meanwhile,” “And,” “For example, say,” “Now,” “Well,” “Okay,” (349)

- Disconfirmation (Dis). Examples: “There are no other options that match your request.” “Not that flight.” “There are no flights with that airline.” “I don’t see anything else.” “No, you must change terminals.” (1,670)
- Filled pause/back-channel (Fill). Examples: “um,” “Hmmm.” “Mmmm.” “uh,” “eh,” “okay.” “Let’s see.” “I’ll see.” (32)

It should be noted that our dialog corpus contains only relatively polite dialogs, and Exclamation-negative utterances are consequently much more limited and not representative of what might be encountered in some other dialog contexts.

2.3. Acoustic measures of prosody

Six acoustic measures of prosody were made based on proprietary 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.

2.4. ToBI Annotation

A portion of the dialog speech act corpus was annotated following the ToBI intonational model [3]. ToBI labeling of AT&T speech corpora was conducted at the Ohio State University Department of Linguistics. Several aspects of this project were previously reported: speed and label assignment of manual ToBI labeling was compared with labeling from automatically predicted labels [4], inter-transcriber reliability was evaluated [5], and tone similarity and inter-transcriber reliability were studied [6].

3. Data Analysis and Results

3.1. Speech Act Differences in Acoustic Measures

A scatter plot of the mean F0 and F0 range of each speech act is shown in Fig. 1. There is wide variation in both mean F0 and pitch range among speech acts. 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 can be divided neatly into two large clusters in Fig. 1 by a negatively sloping imaginary diagonal line from

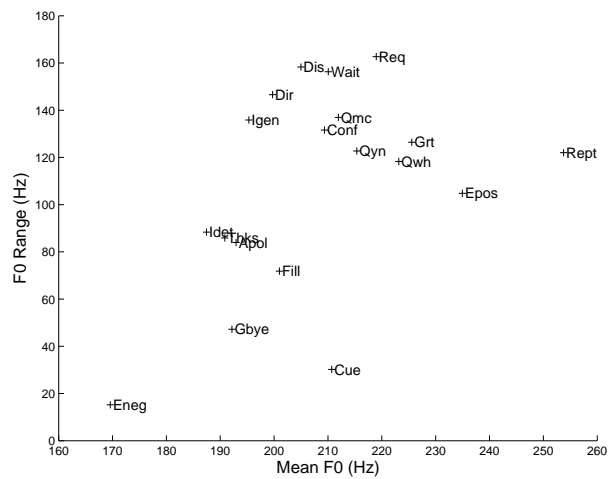


Figure 1: Pitch range and mean F0 of dialog speech acts.

the top of the y-axis to the bottom of the x-axis. One cluster has relatively low mean F0 and F0 ranges and includes: Eneg, Gbye, Cue, Fill, Apol, Thks, and Idet. Speech acts in the second cluster have uniformly higher pitch ranges and tend to have somewhat higher F0 means also. They include Igen, Dir, Dis, Wait, Req, Conf, Qmc, Qyn, Qwh, Grt, Epos, and Rept.

Fig. 2 is a scatter plot of the average phone duration (in ms) and average log power for each speech act. Note that phone duration is plotted on a logarithmic scale because of the extremely wide range of durations observed among speech acts. The average phone duration of Eneg utterances was an extremely long 234 ms. The fastest speaking rate was observed for wh-questions (Qwh) as indicated by an average phone duration of 78 ms. Thus, the speaking rate for wh-questions was three times faster than for Eneg. The long duration of Eneg utterances may be partially explained because of phrase-final lengthening effects on very short exclamations. However, Exclamation-positives (Epos) as well as cue phrases (Cue) were also very brief and subject to the same phrase-final lengthening effects, but have only half the average phone duration of Eneg.

It is interesting to note the large differences in speaking rate (Fig. 2) and F0 range (Fig. 1) between the two Assertive modes: 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.

Log power also differentiated some speech acts from others. Ever the outlier, Eneg utterances had by far the lowest log power at 4.4. Cue and Epos utterances also had relatively low power, at 5.3 and 5.4, respectively. The highest average log power, 6.6, was observed for Disconfirmations (Dis). Log power for most of the remaining speech acts fell within the 6 – 6.5 range.

3.2. Clustering of Speech Acts

A hierarchical cluster analysis was performed from the six acoustic measures of the 19 dialog speech acts. It was of interest to see how speech acts would cluster statistically on the basis of all six acoustic measures of prosody considered together. The results of the cluster analysis are presented in the form of a den-

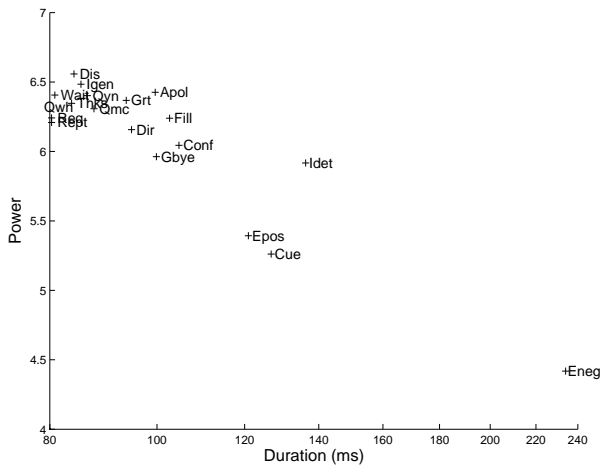


Figure 2: Average phone duration and log power of dialog speech acts.

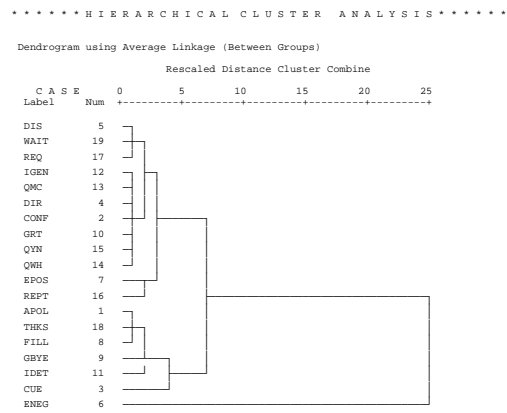


Figure 3: Hierarchical clustering dendrogram of dialog speech acts.

drogram, shown in Fig. 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.

The dendrogram shows, first, that one speech act, Eneg, is extremely different from a large cluster of all the other speech acts. At a second clustering split, the large cluster is then divided into two clusters. The smaller of these clusters is more similar to Eneg than the larger one. The smaller cluster, which we will refer to as cluster A, includes Cue, Idet, Gbye, Fill, Thks, and Apol. There are three additional nodes at which these six speech acts are further split into smaller clusters. Going back up the dendrogram to the second clustering split, the larger cluster of 12 speech acts (cluster B) is subsequently split into two clusters. The smaller cluster consists of Epos and Rept, which soon thereafter form two independent clusters of one speech act each. The larger cluster (C) of 10 speech acts is then split into two clusters. The cluster (D) of three speech acts at the top of the column includes Dis, Wait, and Req. The remaining 7 speech acts, Igen, Qmc, Dir, Conf, Grt, Qyn, and Qwh all cluster together forming what we will refer to as cluster E.

When clusters of the 19 speech acts are related back to their modes of enunciation, several patterns are evident. Note that all the Imperative (Wait, Req, Dir, Rept) and Interrogative (Qmc, Qyn, Qwh) speech acts fall into Cluster B. The two most emotionally positive Affective speech acts, Grt and Epos, are also included in Cluster B. The remaining Affective speech acts, Apol, Thks, Gbye, and Eneg, fall into or below Cluster A. The Assertive mode is split between Cluster B (Igen) and Cluster A (Idet). Of the “Other” (unclassified or ambiguous) category, Disconfirmation (Dis) and Confirmation (Conf), both active speech acts in terms of advancing a directed dialog, are in Cluster B while Filled pause/back-channel (Fill) and Cue phrase (Cue), more passive speech acts that may serve social functions but don’t actively direct the dialog, are in Cluster A.

3.3. ToBI Differences among speech acts

For the purposes of this paper, no large-scale analysis of intonational features and their relation to speech acts will be attempted. However, some examples may serve to indicate that the consideration of speech acts in intonation is a fruitful line of future research and has practical implications for the design of TTS systems.

Occasionally, presumably because of its communicative purpose in the dialog, the intonation pattern of an utterance differed from what would be expected based on a standard syntactic analysis of the sentence. The utterance “What was that?” is an example. In the dialog context, it was classified as the imperative speech act Repeat, since it was clearly intended to elicit a repetition. Based on a traditional syntactic analysis of the text, however, it would be classified as a wh-question, and phrase-final edge tone L-L% would be expected. Although it clearly was not a question that could be answered by yes or no, the actual intonation pattern of this sentence resembled a yes/no question because of the L* nuclear accent on “what” and the H-H% edge tone.

Another example of variation in intonation pattern influenced by speech act status involves multiple choice questions (Qmc). A question such as “Do you prefer a compact car or a standard car?” classified by speech act as Qmc would be classified as a yes/no question by the TTS text analysis module, and a L* H-H% intonation pattern would be predicted. The question “What would you prefer, a compact car or a standard car?” is also classified by speech act as Qmc, but standard text analysis would consider it a wh-question and assign a L-L% final edge tone. Instead, the actual intonation pattern for both utterances was similar: the first option (“a compact car”) formed an intonational phrase with a H-H% edge tone, and the second option in the following phrase (“a standard car”) ended with L-L%. This is admittedly a difficult problem for text analysis, because it depends on whether “or” is intended in its exclusive or inclusive sense. The dialog context, however, disambiguates the situation.

4. Implications of speech acts to TTS

4.1. Providing speech act information to TTS

In spoken dialog systems, used for human-computer dialog, the dialog manager specifies the purpose of an utterance. This goal or purpose is equivalent to a speech or dialog act, although the terminology or categories may differ somewhat from our current usage. A language generation module determines the wording of the utterance, and a speech synthesis system generates audible speech output. In dialog systems, it would be a simple matter to convey the intended speech act to a TTS system designed to use that information at various levels in synthesizing speech.

Other alternatives to providing speech act information to TTS include an analysis of input text to predict the most likely speech act intended or manual text mark-up.

4.2. Use of speech acts by the TTS front end

The front end of a TTS system 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 there are, at least for the speaker studied, systematic differences in pitch, pitch range, phone duration, and power among different speech acts. We have also seen examples, above, where speech act category strongly influences the prosody of an utterance, even to the point of overruling syntactic structure. There seems little doubt that including speech act information along with input text would improve the capability of a TTS front end to assign more appropriate prosody. Without such information, the safest strategy has been to be very conservative with respect to prosodic variation.

4.3. Speech acts and the unit selection TTS back end

The back end of a TTS system accepts the symbolic input of the front end and generates a corresponding speech signal. There are various approaches to speech synthesis, including rule-based formant synthesis, concatenative diphone synthesis, and concatenative unit selection synthesis. We will address only the latter, which is the technique used by our system, AT&T Natural VoicesTMTTS.

Unit selection TTS synthesizes speech by concatenating selected units from a large inventory of several hours of continuous speech recorded from a given speaker. A unit selection algorithm selects a sequence of speech units that best matches the targets (the desired characteristics as determined by the TTS front end) and also that join together most smoothly. Typically, little signal processing is performed on the resulting speech signal, so there is minimal distortion. Unit selection synthesis results in more natural sounding speech than the other synthesis techniques, but its prosody is more difficult to control. Our hypothesis is that the use of speech acts will provide more prosodic control of TTS and also that TTS control via speech acts will be at a more accessible level than that requiring specialized linguistic knowledge.

We have implemented a prototype dialog TTS system and inventory that includes the speech act of the original recorded utterance as one of the features upon which unit selection is based. If the original speech act of a unit and the speech act to be synthesized do not match, the unit selection algorithm penalizes the use of that unit. Our approach of redesigning the TTS back end was taken because voice quality (as it relates to phonation and laryngeal setting and also gestures such as smil-

ing), although not acoustically measured in this study, carries another dimension of expressiveness apart from prosody [2] that we wanted to capture. Although our work on the TTS prototype is in its early stages and refinement of the system remains to be done, the results sound promising and will be demonstrated.

5. 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 and their higher level modes of enunciation form meaningful groups when hierarchically clustered on the basis of their acoustic measures.

Examples demonstrate that speech act category can strongly influence intonational contours, even overriding the contour expected from a standard analysis of their syntactic structure.

There are several practical implications of the study of speech acts for TTS. Dialog systems generate speech act information and can provide it to TTS. Inclusion of speech act category can improve the assignment of prosody in the TTS front end. A speech inventory and unit selection TTS system that includes speech acts as features has additional potential to control TTS prosody and to improve naturalness of TTS in dialogs.

6. References

- [1] Beutnagel, M.; Conkie, A.; Schroeter, J.; Stylianou, Y.; Syrdal, A., 1999. The AT&T Next-Gen TTS system. *Proc. Joint Meeting of ASA, EAA, and DEGA*. Germany: Berlin, SASCA_4, <http://www.research.att.com/projects/tts/pubs.html>.
- [2] Bulut, M.; Narayanan, S.; Syrdal, A. K., 2002. Expressive speech synthesis using a concatenative synthesizer. *International Conference on Spoken Language Processing*. Colorado: Denver, 1265-1268.
- [3] Silverman, K.; Beckman, M.; Pierrehumbert, J.; Ostendorf, M.; Wightman, C.; Price, P.; Hirschberg, J., 1992. ToBI: A standard scheme for labeling prosody. *International Conference on Spoken Language Processing*. Canada: Banff, 867-879.
- [4] Syrdal, A. K.; Hirschberg, J.; McGory, J.; Beckman, M., 2001. Automatic ToBI prediction and alignment to speed manual labeling of prosody. *Speech Communication*, 33, 135-151.
- [5] Syrdal, A. K.; McGory, J., 2000. Inter-transcriber reliability of ToBI prosodic labeling. *International Conference on Spoken Language Processing*. China: Beijing, III, 235-238.
- [6] Herman, R.; McGory, J. T., 2002. The conceptual similarity of intonational tones and its effects on intertranscriber reliability. *Language and Speech*, 45(1), 1-36.