

Using Dialogue Representations for Concept-to-Speech Generation

Christine H. Nakatani
Jennifer Chu-Carroll

Bell Laboratories, Lucent Technologies
600 Mountain Avenue
Murray Hill, NJ 07974 USA

{chn|jenc}@research.bell-labs.com

Abstract

We present an implemented concept-to-speech (CTS) system that offers original proposals for certain couplings of dialogue computation with prosodic computation. Specifically, the semantic interpretation, task modeling and dialogue strategy modules in a working spoken dialogue system are used to generate prosodic features to better convey the meaning of system replies. The new CTS system embodies and extends theoretical work on intonational meaning in a more general, robust and rigorous way than earlier approaches, by reflecting compositional aspects of both dialogue and intonation interpretation in an original computational framework for prosodic generation.

1 Introduction

Conversational systems that use speech as the input and output modality are often realized by architectures that decouple speech processing components from language processing components. In this paper, we show how speech generation can be more closely coupled with the dialogue manager of a working mixed-initiative spoken dialogue system. In particular, we use representations from the semantic interpretation, task model and dialogue strategy modules to better communicate the meaning of system replies through prosodically appropriate synthetic speech.

While dialogue prosody has been a topic of much study, our implemented concept-to-speech (CTS) system offers original proposals for specific couplings of dialogue computation with prosodic computation. Further, it embodies and extends theoretical work on intonational meaning in a more general, robust and rigorous way than earlier CTS systems, in an architecture that reflects compositional aspects of dialogue and intonation interpretation.

2 Theoretical Foundations

In this work, we implement and extend the compositional theory of intonational meaning proposed by Pierrehumbert and Hirschberg (1986; 1990), who sought to identify correspondences between the

Grosz and Sidner (1986) computational model of discourse interpretation and Pierrehumbert's prosodic grammar for American English (1980).

In the present work, certain aspects of the original theories are modified and adapted to the architecture of the dialogue system in which the CTS component is embedded. Below, we present the important fundamental definitions and principles of intonation underlying our CTS system.

2.1 Intonational System

In our CTS system, the prosodic elements that are computed are based on the intonational system of Pierrehumbert (1980), who defined a formal language for describing American English intonation using the following regular grammar:

$$\begin{aligned} \text{Inton Phrase} &\rightarrow (\text{Interm Phrase})^+ \text{Bndry Tone} \\ \text{Interm Phrase} &\rightarrow (\text{Pitch Acc})^+ \text{Phrase Acc} \end{aligned}$$

Major phrases, or *intonational phrases*, are made up of one or more minor phrases, or *intermediate phrases*. Melodic movements in intermediate and intonational phrases are in turn expressed by three kinds of tonal elements. These include six *pitch accents*: a low pitch excursion (L^*), a high pitch excursion (H^*), or a combination of both low and high excursions (L^*+H , $L+H^*$, H^*+L , $H+L^*$); two *phrase accents*: a high (H -) or low (L -) tonal target that guides the interpolation of the melodic contour from final pitch accent to intermediate phrase ending; and two *boundary tones*: a high ($H\%$) or low ($L\%$) tonal target that guides interpolation from phrase accent to intonational phrase ending.

2.2 Intonational Meaning

Theoretical work on intonational meaning has attempted to relate the grammatical elements of Pierrehumbert's system — pitch accent, phrase accent and boundary tone, to interpretive processes at different levels of discourse and dialogue structure. Hirschberg and Pierrehumbert (1986) conjectured that the absence or presence of accentuation conveys discourse focus status, while the tonal properties of the accent itself (i.e. pitch accent type) convey semantic focus information.

MIMIC: hello this is mimic the movie information system
how can I help you
User: where in hoboken is october sky playing
MIMIC: october sky is playing at hoboken cinema in hoboken
can I help you with anything else
User: when is it playing there
MIMIC: october sky is playing at hoboken cinema in hoboken at 3:45pm, 5:50pm, 7:10pm, and 10pm
can i help you with anything else

Figure 1: A MIMIC dialogue.

In later work, pitch accent type was said to express whether the accented information was intended by the speaker to be “predicated” or not by the hearer (Pierrehumbert and Hirschberg, 1990). Non-predicated information was said to bear low-star accentuation (L^* , L^*+H , $H+L^*$), while predicated information would be marked by high-star accents (H^* , $L+H^*$, H^*+L). The theory further stated that L^*+H conveys uncertainty or lack of speaker commitment to the expressed propositional content, while $L+H^*$ marks correction or contrast. The complex accent, H^*+L , was said to convey that an inference path was required to support the predication; usage of $H+L^*$ similarly was said to imply an inference path, but did not suggest a predication of a mutual belief. Finally, phrase accents and boundary tones were said to reflect aspects of discourse structure.

3 Systems Foundations

Our task is to improve the communicative competence of a spoken dialogue agent, by making recourse to our knowledge of intonational meaning, dialogue processing and relations between the two. Of course, a worthwhile CTS system must also outperform out-of-the-box text-to-speech (TTS) systems that may determine prosodic mark-up in linguistically sophisticated ways. As in (Nakatani, 1998), we take the prosodic output of an advanced research system that implements the Pierrehumbert theory of intonation, namely the Bell Labs TTS system, as our baseline experimental system to be enhanced by CTS algorithms. We embed the CTS system in MIMIC, a working spoken dialogue system representing state-of-the-art dialogue management practices, to develop CTS algorithms that can be eventually realistically evaluated using task-based performance metrics.

3.1 Dialogue System: Mixed-Initiative Movie Information Consultant (MIMIC)

The dialogue system whose baseline speech generation capabilities we enhance is the Mixed-Initiative Movie Information Consultant (MIMIC) (Chu-Carroll, 2000). MIMIC provides movie list-

ing information involving knowledge about towns, theaters, movies and showtimes, as demonstrated in Figure 1. MIMIC currently utilizes template-driven text generation, and passes on text strings to a stand-alone TTS system. In the version of MIMIC enhanced with concept-to-speech capabilities, MIMIC-CTS, contextual knowledge is used to modify the prosodic features of the slot and filler material in the templates; we are currently integrating the algorithms in MIMIC-CTS with a grammar-driven generation system. Further details of MIMIC are presented in the relevant sections below, but see (Chu-Carroll, 2000) for a complete overview.

3.2 TTS: The Bell Labs System

For default prosodic processing and speech synthesis realization, we use a research version of the Bell Labs TTS System, circa 1992 (Sproat, 1997), that generates intonational contours based on Pierrehumbert’s intonation theory (1980), as described in (Pierrehumbert, 1981). Of relevance is the fact that various pitch accent types, phrase accent and boundary tones in Pierrehumbert’s theory are directly implemented in this system, so that by generating a Pierrehumbert-style prosodic transcription, the work of the CTS system is done. More precisely, MIMIC-CTS computes prosodic annotations that override the default prosodic processing that is performed by the Bell Labs TTS system.

To our knowledge, the intonation component of the Bell Labs TTS system utilizes more linguistic knowledge to compute prosodic annotations than any other unrestricted TTS system, so it is reasonable to assume that improvements upon it are meaningful in practice as well as in theory.

4 MIMIC’s Concept-to-Speech Component (MIMIC-CTS)

In MIMIC-CTS, the MIMIC dialogue system is enhanced with a CTS component to better communicate the meaning of system replies through contextually conditioned prosodic features. MIMIC-CTS makes use of three distinct levels of dialogue representations to convey meaning through intonation. MIMIC’s *semantic representations* allow MIMIC-CTS to decide *which* information to prosodically

Task Specification Status	Information Status	Pitch Accent
Required (Y)	KEY	L+H*
Optional (-)	INFERRABLE/OLD	L*+H/L*
Not allowed (N)	HEARER-NEW	H*

Table 1: Highlighting relevance of information based on task model (and discourse history).

User: where in montclair is analyze this playing
MIMIC: **analyze this is playing at wellmont theatre and clearviews screening zone in montclair**

ANALYZE	THIS	is	PLAYING	at	WELLMONT	THEATER	and
L+H*	L+H*	-	L*	-	H*	H* L-H%	-
	CLEARVIEWS	SCREENING	ZONE	in	MONTCLAIR		
	H*	H*	H* L-H%	-	L+H*	L-L%	

Figure 4: Above, dialogue excerpt of MIMIC performing a *where* task. Below, the modified version of the bold-faced reply string, generated by MIMIC-CTS.

user, but is derived by MIMIC's limited inference engine that seeks to instantiate as many attribute values as possible. For instance, a theater name may be inferred given a town name, if there is only one theater in the given town. OLD information is inherited from the discourse history, based on updating rules relying on confidence scores for attribute values. HEARER-NEW information (c.f. (Prince, 1988)) is that which is requested by the user, and constitutes the only new information on the scale. But note that KEY information, while given, is still clearly in discourse focus, along with HEARER-NEW information.

The next step is to map the information statuses, ordered from given to new, to a scale of pitch accent, or accent melodies, ordered from given to new as follows:

L*	L*+H	L+H*	H*
[given]			[new]

Table 1 summarizes this original mapping of information statuses to pitch accent melodies, and Figure 4 illustrates the use of this mapping in an example. It obeys the general principle of Pierrehumbert and Hirschberg's work, that low tonality signifies discourse givenness and high tonality signifies discourse newness, but extends this principle beyond its vague definition in terms of predication of mutual beliefs. Instead, the principle is operationalized here in a practically motivated manner that is consistent with and perhaps illuminating of the theory.

4.3 Assigning "Dialogue Prosody" using Dialogue Strategies

As in earlier CTS systems, special logico-semantic relations, such as contrast or correction, are effectively conveyed in MIMIC-CTS by prosodic cues. In MIMIC-CTS, however, these situations are not stipulated in an *ad hoc* manner, but can be determined to a large degree by MIMIC's dialogue strategy selection process that identifies appropriate dialogue acts to realize a dialogue goal.³

For example, the dialogue act ANSWER may be selected to achieve the dialogue goal of providing an answer to a successful user query, while the dialogue act NOTIFYFAILURE may be performed to achieve the dialogue goal of providing an answer in situations where no movie listing in the database matches the user query. The template associated with the dialogue act, NOTIFYFAILURE, when compared with that for ANSWER, contains an additional negative auxiliary associated with the key attribute responsible for the query failure, in an utterance conveying a contradiction in beliefs between the user and system (namely, the presupposition on the part of the user that the query can be satisfied).

Theoretical work on intonational interpretation leads us to prosodically mark the negative auxiliary, as well as the associated focus position (Rooth, 1985). We choose to mark the negative auxiliary *not* with the L+H* pitch accent to convey correction, while marking the material in the associated focus position with the L*+H pitch accent to convey (the

³Importantly, MIMIC's adaptive dialogue strategy selection algorithm takes into account the outcome of an initiative tracking module that we do not discuss here (see (Chu-Carroll, 2000)).

User:	where is the corruptor playing in cranford					
MIMIC:	the corruptor is not playing in cranford the corruptor is playing at lincoln cinemas in arlington					
THE	CORRUPTOR	is	NOT	playing	in	CRANFORD
L+H*	L+H*	-	L+H*	!H*	-	L*+H L-H%

Figure 5: Above, dialogue excerpt of MIMIC performing a NOTIFYFAILURE dialogue act. Below, the modified version of the bold-faced reply string, generated by MIMIC-CTS. Note the diacritic “!” denotes a *downstepped* accent (see (Pierrehumbert, 1980)).

system’s) lack of commitment to the (user’s) presupposition at hand. Finally, the NOTIFYFAILURE dialogue act is conveyed by assigning the so-called rise-fall-rise contradiction contour, L*+H L-H%, to the utterance at large (c.f. (Hirschberg and Ward, 1991)). An example generated by MIMIC-CTS appears in Figure 5. Note that pitch accent types for the remaining attribute values are assigned using the task model, as described in section 4.2. Thus in Figure 5, the movie title is treated as KEY information, marked by the L+H* pitch accent.

MIMIC-CTS contains additional prosodic rules for logical connectives, and clarification and confirmation subdialogues.

5 Related Work

Although a number of earlier CTS systems have captured linguistic phenomena that we address in our work, the computation of prosody from dialogue representations is often not as rigorous, detailed or complete as in MIMIC-CTS. Further, while several systems use given/new information status to decide whether to accent or deaccent a lexical item, no system has directly implemented general rules for pitch accent type assignment. Together, MIMIC-CTS’s computation of accentuation, pitch accent type and dialogue prosody constitutes the most general and complete implementation of a compositional theory of intonational meaning in a CTS system to date.

Nevertheless, elements of a handful of previous CTS systems support the approaches taken in MIMIC-CTS toward conveying semantic, task and dialogue level meaning. For example, the Direction Assistant system (Davis and Hirschberg, 1988) mapped a hand-crafted route grammar to a discourse structure for generated directions. The discourse structure determined accentuation, with deaccenting of discourse-old entities realized (by lexically identical morphs) in the current or previous discourse segment. Other material was assigned accentuation based on lexical category information, with the exception that certain contrastive cases of accenting, such as *left* versus *right*, were stipulated for the domain.

Accent assignment in the SUNDIAL travel infor-

mation system (House and Youd, 1990) also relied on discourse and task models. Mutually known entities, said to be in *negative focus*, were deaccented; entities in the current task space, in *referring focus*, received (possibly contrastive) accenting; and entities of the same type as a previously mentioned object, were classified as in either *referring* or *emphatic focus*, depending on the dialogue act. In the cases of corrective situations or repeated system-initiated queries, the contrasting or corrective items were emphatically accented.

The BRIDGE project on speech generation (Zacharski et al., 1992) identified four main factors affecting accentability: linear order, lexical category, semantic weight and givenness. In related work (Monaghan, 1994), word accentability was quantitatively scored by hand-crafted rules based on information status, semantic focus and word class. The givenness hierarchy of Gundel and colleagues (1989), which associates lexical forms of expression with information statuses, was divided into four intervals, with scores assigned to each. A binary semantic focus score was based on whether the word occurred in the *topic* or *comment* of a sentence. Finally, lexical categories determined word class scores. These scores were combined, and metrical phonological rules then referred to final accentability scores to assign a final accenting pattern.

To summarize, all of the above CTS systems employ either hand-crafted or heuristic techniques for representing semantic and discourse focus information. Further, only SUNDIAL makes use of dialogue acts.

6 Conclusion and Future Work

We are presently carrying out evaluations of MIMIC-CTS. An initial corpus-based analysis compares the prosodic annotations assigned to three actual MIMIC dialogues, which were previously collected during an overall system evaluation (Chu-Carroll and Nickerson, 2000). The corpus of dialogues is made up of 37 system/user turns, including 40 system-generated sentences. Three versions of the MIMIC dialogues are being analysed, with prosodic features arising from three differ-

ent sources: MIMIC-CTS, MIMIC operating with default Bell Labs TTS, and a professional voice talent who read the dialogue scripts in context. This corpus-based assessment — comparing the prosody of CTS-generated, TTS-generated, and human speech, will enable more domain-dependent tuning of the MIMIC-CTS algorithms, as well as the refinement of general prosodic patterns for linguistic structures, such as lists and conjunctive phrases. Ultimately, the value of MIMIC-CTS must be measured based on its contribution to overall task performance by real MIMIC users. Such a study is under design, following (Chu-Carroll and Nickerson, 2000).

In conclusion, we have shown how prosodic computation can be conditioned on various dialogue representations, for robust and domain-independent CTS synthesis. While some rules for prosody assignment depend on the task model, others must be tied closely to the particular choices of content in the replies, at the level of dialogue goals and dialogue acts. At this level as well, however, linguistic principles of intonation interpretation can be applied to determine the mappings. In sum, the lesson learned is that a unitary notion of “concept” from which we generate a unitary prosodic structure, does not apply to state-of-the-art spoken dialogue generation. Instead, the representation of dialogue meaning in experimental architectures, such as MIMIC’s, is compositional to some degree, and we take advantage of this fact to implement a compositional theory of intonational meaning in a new concept-to-speech system, MIMIC-CTS.

References

- Jennifer Chu-Carroll and Bob Carpenter. 1999. Vector-based natural language call routing. *Computational Linguistics*, 25(3):361–388.
- Jennifer Chu-Carroll and Jill S. Nickerson. 2000. Evaluating automatic dialogue strategy adaptation for a spoken dialogue system. In *Proceedings of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, Seattle.
- Jennifer Chu-Carroll. 2000. Mimic: an adaptive mixed initiative spoken dialogue system for information queries. In *Proceedings of the 6th Conference on Applied Natural Language Processing*, Seattle.
- J. R. Davis and J. Hirschberg. 1988. Assigning intonational features in synthesized spoken directions. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 187–193, Buffalo.
- Barbara Grosz and Candace Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.
- J. Gundel, N. Hedberg, and R. Zacharski. 1989. Givenness, implicature and demonstrative expressions in English discourse. In *Proceedings of CLS-25, Parasession on Language in Context*, pages 89–103. Chicago Linguistics Society.
- Julia Hirschberg and Janet Pierrehumbert. 1986. The intonational structuring of discourse. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, New York.
- J. Hirschberg and G. Ward. 1991. The influence of pitch range, duration, amplitude, and spectral features on the interpretation of $l^*+h\ l\ h\%$. *Journal of Phonetics*.
- Jill House and Nick Youd. 1990. Contextually appropriate intonation in speech synthesis. In *Proceedings of the European Speech Communication Association Workshop on Speech Synthesis*, pages 185–188, Autrans.
- A. I. C. Monaghan. 1994. Intonation accent placement in a concept-to-dialogue system. In *Proceedings of the ESCA/IEEE Workshop on Speech Synthesis*, pages 171–174, New Paltz, NY.
- C. H. Nakatani. 1998. Constituent-based accent prediction. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, Montreal.
- J. Pierrehumbert and J. Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In *Intentions in Communication*. MIT Press, Cambridge, MA.
- Janet Pierrehumbert. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, Massachusetts Institute of Technology, September. Distributed by the Indiana University Linguistics Club.
- J. Pierrehumbert. 1981. Synthesising intonation. *Journal of the Acoustical Society of America*, 70(4):985–995.
- Ellen Prince. 1988. The ZPG letter: subjects, definiteness, and information status. In S. Thompson and W. Mann, editors, *Discourse Description: Diverse Analyses of a Fund Raising Text*. Elsevier Science Publishers, Amsterdam.
- Mats Rooth. 1985. *Association with Focus*. Ph.D. thesis, University of Massachusetts, Amherst MA.
- Richard Sproat, editor. 1997. *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Kluwer Academic, Boston.
- Ron Zacharski, A. I. C. Monaghan, D. R. Ladd, and Judy Delin. 1992. BRIDGE: Basic research on intonation for dialogue generation. Technical report, University of Edinburgh.