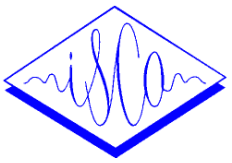

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

COUPLING DIALOGUE AND PROSODY COMPUTATION IN SPOKEN DIALOGUE GENERATION

Christine H. Nakatani[d]

chn@nuance.com Dialog R&D, Nuance Communications
1380 Willow Road, Menlo Park, CA 94025 USA

Jennifer Chu-Carroll

jencc@research.bell-labs.com Bell Labs, Lucent Technologies
600 Mountain Ave, Murray Hill, NJ 07974 USA

ABSTRACT

We introduce a concept-to-speech (CTS) system that generates prosodic structure compositionally, in a spoken dialogue agent architecture. Representations from the semantic interpretation, task modeling, and dialogue strategy selection components drive the computation of accentuation, pitch accent type selection, and choice of melodic contour, respectively. These principled couplings of dialogue and prosody computation extend both the theory and practice of concept-to-speech generation.

1. OVERVIEW

A perennial problem in concept-to-speech (CTS) generation has been the proper definition of "concept" representations, from which prosodic features may be computed. We argue that there is no unitary concept of dialogue meaning, in theory or practice, from which a unitary prosodic structure can be computed. Instead, prosodic structure must be built compositionally, by defining theoretically rigorous but robust couplings of dialogue and prosodic computation. In this paper, we identify three general components of CTS generation, coupling semantic interpretation and accentuation, task modeling and pitch accent type selection, and dialogue strategy selection and choice of melodic contour. These couplings of dialogue and prosodic computation arise from their shared role in the modeling of discourse focusing, semantic focusing, and discourse intentions, respectively.

Although we illustrate our compositional approach to CTS generation in a working template-driven generation system embedding advanced text-to-speech (TTS) technology within an innovative modular dialogue agent architecture, the underlying theoretical insights should allow the principled generalization of this approach to other generation techniques, dialogue architectures, and TTS systems.

1.1. Theoretical Foundations

In our CTS system, we implement and extend the theory of intonational meaning proposed by Pierrehumbert and Hirschberg (1990), who sought to identify correspondences between the Grosz and Sidner (1986) computational model of discourse interpretation and Pierrehumbert's (1980) prosodic grammar for American English. Our CTS system represents the most complete application of these theories of intonation and discourse structure in a spoken dialogue generation system.

Intonational System. Our CTS system computes prosodic structure based on Pierrehumbert's intonational system, which is described by the following regular grammar:

$$\text{IntermediatePhrase} \rightarrow (\text{PitchAccent})^{+}\ \text{PhraseAccent}$$
$$\text{IntonationalPhrase} \rightarrow (\text{IntermediatePhrase})^{+}\ \text{BoundaryTone}$$

Intermediate or minor phrases consist of one or more *pitch accents* or accent melodies (associated with word units), followed by a high (H-) or low (L-) *phrase accent* that guides interpolation of the melodic contour from final pitch accent to intermediate phrase ending. Pitch accents may consist of 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*), in which the starred tone is generally perceptually more dominant. *Intonational* or major phrases, consist of one or more intermediate or minor phrases plus a *boundary tone*, or final high (H%) or low (L%) tonal target that guides interpolation from phrase accent to intonational phrase ending.

Intonational Meaning. Theoretical work on intonational meaning has attempted to relate the prosodic elements of

Pierrehumbert's intonational grammar, to interpretive processes at different levels of discourse and dialogue structure. Pierrehumbert and Hirschberg (1990) conjectured that the absence or presence of accentuation conveys "the relative salience of accented items in the discourse", while the tonal properties of the accent itself (i.e. pitch accent type) convey "other sorts of information status" (p. 286). More precisely, pitch accent type was said to express whether the accented information was intended by the speaker to be "predicated" or not by the hearer, in a mutual belief framework of discourse modeling. 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). Finally, phrase accents and boundary tones were said to reflect discourse segmental structure (Grosz and Sidner, 1986).

1.2. Systems Foundations

The goal of CTS research is to improve the communicative competence of a spoken dialogue agent, by the principled use of prosody to convey linguistic meaning. Of course, a worthwhile CTS system must also outperform out-of-the-box TTS systems that may determine prosodic mark-up in linguistically sophisticated ways. As in related previous work (Nakatani, 1998), we take the prosodic output of an advanced research system, 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.

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 listing information, as demonstrated in Figure 1.

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.

MIMIC embodies an innovative mixed-initiative, modular spoken dialogue agent architecture, whose full details lie beyond the scope of this paper. It currently utilizes template-driven text generation, and passes on text strings to a stand-alone TTS system. In the version of MIMIC enhanced with CTS, MIMIC-CTS, contextual knowledge is used to modify the prosodic features of the slot and filler material in the templates. The principles of processing in MIMIC-CTS, however, may be applied across a range of generation and component processing techniques. What critically enables the implementation of MIMIC-CTS is MIMIC's modularization of the dialogue components that perform semantic interpretation, task modeling, and dialogue strategy selection.

TTS: The Bell Labs System. For default prosodic processing and speech synthesis realization, we use a research version of the Bell Labs TTS System (Sproat, 1997). MIMIC-CTS computes prosodic annotations of template-generated text strings that override default processing. Prosodic modeling in the Bell Labs research system is based on Pierrehumbert's theory of intonation, as described in (Pierrehumbert, 1981).
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.

2. CONCEPT-TO-SPEECH SYSTEM

In MIMIC-CTS, three different dialogue processing components drive the computation of prosodic features. MIMIC's semantic interpretation module allows MIMIC-CTS to identify *which* information to prosodically highlight. MIMIC's task model in turn determines *how* to prosodically highlight selected information. Finally, MIMIC's dialogue strategy selection process informs various choices in the assignment of stylized melodic contours that convey logico-semantic and pragmatic aspects of meaning.

2.1. Highlighting Information Using Semantic Representations

MIMIC uses a statistically-driven semantic interpretation engine to identify values for a fixed list of attributes that serve as keys in queries, such as movie titles, and theater and town names. Both attribute names and attribute values represent information in *discourse focus*, or salient with respect to the task at hand. In MIMIC-CTS, attribute names and values that occur in

generation templates are semantically typed by the attribute name. Typed information is then prosodically highlighted in the following way:

1. All lexical items realizing attribute values are accented.
2. Attribute values are synthesized at a slower speaking rate.
3. Attribute values are set off by phrase boundaries.
4. Attribute names are always accented.

Even though these modifications are entirely rule-based, given a typing system, highlighting attribute names in particular is often overlooked in spoken dialogue generation. For example, while the default TTS system would say, "What movie would you LIKE?" MIMIC-CTS would more helpfully accent the attribute name *movie*, and say, "What MOVIE would you like?"

This approach to accentuation importantly decouples the problem of making *accent/deaccent* decisions from the problem of assigning accent melody, or pitch accent type. MIMIC-CTS represents the first general implementation of this theoretically motivated decomposition of the "accent assignment" problem in a CTS system. We argue that this modularization of prosodic processing is justified by the insight that *discourse* focusing versus *semantic* focusing aspects of meaning are realized by distinct, yet intimately related, elements of prosodic structure. As described below, this computational decoupling of the accentuation component from the accent realization component, enables a novel pitch accent type assignment algorithm.

2.2. Conveying Information Status Using Task Modeling

Next, MIMIC's task model determines how to prosodically highlight selected information, based on the pragmatic properties of the system reply. The task model, shown in Table 1, defines which attributes are *required*, *not allowed*, or *optional*, to license a database query for a closed set of information-giving tasks. For example, to elicit movieshowtimes, the user must provide both movie and theater values as required by the *when* task.

Task	**Movie**	**Theater**	**Town**
What	*Not allowed*	*Required*	*Optional*
Where	*Required*	*Not allowed*	*Required*
When	*Required*	*Required*	*Optional*

Table 1: Task Specifications for MIMIC.

To better convey the structure of the task model, which is learned by the user through interaction with the system, we annotate the generation templates that convey query results to the user with the information status of each attribute value occurring in the template. Information status is determined by analysis of the semantic focusing properties of the task specification statuses, in the context of the user query. Then, when MIMIC-CTS generates a system reply, each attribute value is assigned a particular melody or pitch accent type that conveys its information status, as shown in Table 2.

Task Specifica-Tion Status	**Information Status**	**Pitch Accent Type**
Not allowed	HEARER-NEW	H*
Required	KEY	L+H*
Optional	INFERRABLE	L*+H
Optional	OLD	L*

Table 2: Mapping of task specification status to the pitch accent type that highlights the relevance of attribute value information in the context of the user query.

Information that the user intends to elicit from the system is marked *not allowed* in the task specification and is assigned the HEARER-NEW information status (Prince, 1988). HEARER-NEW information corresponds to the semantic focus proper, and is marked by the most prominent accent type, H*. *Required* information that is necessary to formulate a valid database query is considered KEY. KEY information has been provided by the user, but is conveyed in the system response to both implicitly confirm the user query and supply context for interpretation. We assign KEY information the next most prominent accent type, L+H*. Less critical to user comprehension of the system reply is INFERRABLE information, which is derived by MIMIC's limited inference engine that instantiates as many attribute values as possible, by inferring a town given a theater name for instance. INFERRABLE information receives the somewhat prominent L*+H accent. Information that is inherited from the user discourse history is OLD, and this information status is conveyed by the least prominent accent type, L*. For example, anaphoric pronouns in MIMIC are resolved to discourse history items during interpretation; in generation, the full form of referring expression is appropriately marked by L* (Nakatani, 1997).

This original scheme for assigning pitch accent type, illustrated by example in Figure 2, operationalizes the general principle of Pierrehumbert and Hirschberg's theory of intonational meaning, that low tonality signifies discourse givenness and high tonality signifies

discourse newness. At the same time, it is more general and robust than the various approaches that drive pitch accent prediction off of a unitary "grammatical" representation, such as a "theme/rheme" utterance parse or a "F(ocus)-marked" phrase structure. Many systems based on these types of concept representations distinguish only two levels of information status for referents: thematic and rhematic, or focus and non-focus. They also often impose empirically unfounded formal constituency constraints on semantic focusing representations. As stated, we do not believe that prosodic structure derives from any such unitary representation of meaning. Our approach is also interestingly more general than stochastic approaches to date, and in so far as we have identified the proper representations involved and the correct computational couplings, this fact should not be surprising. Typically, stochastic approaches conflate accentuation and pitch accent type selection, utilize impoverished semantic focusing and task models, and in practice, often suffer from sparse data problems.

Dialogue Excerpt

User: Where in Montclair is Casablanca playing?
MIMIC-CTS: Casablanca is playing at Wellmont Theater and Clearviews Screening Zone in Montclair.

MIMIC-CTS Template Annotations

[*movie*, KEY **Casablanca L+H***] is – playing L* at – [*theater*, HEARER-NEW **Wellmont H* Theater H***] L-H% and – [*theater*, HEARER-NEW **Clearviews H* Screening H* Zone H***] in – [*town*, KEY **Montclair L+H***] L-L%

Figure 2: Excerpt of MIMIC-CTS performing an answer to a *where* task.

2.3. Assigning "Dialogue Prosody" Using Dialogue Strategies

Finally, MIMIC's dialogue strategy selection process informs various choices in prosodic contour and accenting that convey logico-semantic aspects of meaning. MIMIC-CTS currently implements special dialogue prosody mark-up for contradictions, logical connectives, and clarification and confirmation subdialogues. Unlike most CTS systems, however, these cases 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.

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*, expresses a contradiction in beliefs between the user and system (namely, the presupposition on the part of the user that the query can be satisfied). To convey contradiction intonationally, we mark the negative auxiliary, as well as the associated focus position (Rooth, 1985). The former is marked with the L+H* pitch accent to convey correction, while the latter is marked by L*+H, which expresses uncertainty in the context of the overall contradiction contour, the so-called rise-fall-rise, L*+H L-H (c.f. Hirschberg and Ward, 1991). An example generated by MIMIC-CTS appears in Figure 3.

Dialogue Excerpt

User: Where is Adam's Rib playing in Summit?
MIMIC-CTS: Adam's Rib is not playing in Summit. Adam's Rib is playing at Rialto Theatre in Westfield.

MIMIC-CTS Prosodic Annotations

Adam's L+H* Rib L+H* is – **not L+H*** playing !H* in - **Summit L*+H L-H%** Adam's L+H* Rib L+H* is – playing L* at – Rialto H* Theater H* in – Westfield L+H* L-L%

Figure 3: Excerpt of MIMIC-CTS performing a *NotifyFailure* dialogue act, followed by a cooperative, system-initiated answer to a *where* task.[1]

3. CONCLUSION

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. For example, while several systems use given/new information status to decide whether to accent or deaccent a lexical item (Davis and Hirschberg, 1988; Mohaghan, 1994), no system has directly implemented general rules for pitch accent type assignment. Also, MIMIC-CTS captures the linguistic insights of earlier hand-crafted CTS systems that utilize dialogue act information (House and Youd, 1990), but in a principled and robust manner. 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.

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 semantic and task models, 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.

4. REFERENCES

J. Chu-Carroll, 2000. MIMIC: an adaptive mixed-initiative spoken dialogue system for information queries. ANLP-00, Seattle.

J. Davis and J. Hirschberg, 1988. Assigning intonational features in synthesized spoken directions. ACL-88, Buffalo.

B. J. Grosz and C. Sidner, 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3).

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*.

J. House and N. Youd, 1990. Contextually appropriate intonation in speech synthesis. ESCA Workshop on Speech Synthesis. Autrans.

A. I. Monaghan, 1994. Intonation accent placement in concept-to-dialogue system. ESCA/IEEE Workshop on Speech Synthesis. New Paltz.

C. H. Nakatani, 1997. Discourse structural constraints on accent in spontaneous narrative. In J.P.H. van Santen, R. Sproat, J. Olive and J. Hirschberg (eds.), *Progress in Speech Synthesis*. Springer Verlag, New York.

C. H. Nakatani, 1998. Constituent-based accent prediction. ACL-98, Montreal.

J. Pierrehumbert and J. Hirschberg, 1990. The meaning of intonational contours in the interpretation of discourse. In
P. Cohen, J. Morgan and M. Pollack (eds.), *Intentions in Communciation*. MIT Press, Cambridge, MA.

J. Pierrehumbert 1980. The phonology and phonetics of English intonation. Ph.D. Thesis. MIT, Cambridge, MA.

J. Pierrehumbert, 1981. Synthesising intonation. *Journal of the Acoustical Society of America*, 70(4).

E. Prince, 1988. The ZPG letter: subjects, definiteness, and information status. In S. Thompson and W. Mann (eds.), *Discourse Description*. Elsevier Science Publishers, Amsterdam.

M. Rooth, 1985. *Association with Focus*. Ph.D. Thesis. University of Massachusetts, Amherst.

R. Sproat (ed.), 1997. *Multilingual Text-to-Speech Synthesis*. Kluwer Academic, Boston.

[δ] The first author carried out this research at Bell Labs, Lucent Technologies, Murray Hill, NJ, USA.

[1] In this case, MIMIC-CTS is a bit over-robust. The annotations on the second utterance match those of a *where* task; the town value is treated as *required*, KEY information, whereas in reality it is HEARER-NEW, and also contrastive. We can ameliorate this lack of sensitivity to dialogue context by integrating the novel *initiative-tracking* component of MIMIC with MIMIC-CTS, and considering for each *Answer* dialogue act that the system performs, whether the user or the system holds the initiative. In MIMIC, *cooperative responses*, such as the second utterance in Figure 3, are generated only when the system holds the initiative. See (Chu-Carroll, 2000) for discussion of MIMIC's initiative modeling.