Research Statement

Gina-Anne Levow

Speech and language processing have made dramatic strides in the last two decades. Through the use of machine learning techniques and other stochastic approaches ranging from the Hidden Markov Models which revolutionized speech recognition to stochastic context-free grammar induction for parsing, language processing has become more robust and of broader coverage. Speech recognition on relatively constrained tasks has reached high levels of accuracy for speaker-independent recognition; airline status information can be recognized with accuracies well over 95% allowing widespread commercial deployment. Speech by broadcast news anchors can be recognized with less that 10% word error rate, allowing information retrieval at levels comparable to manual transcriptions using term vector-based techniques. However, a recent Computerworld article still ranks speech recognition as one of the "21 Greatest Technology Flops".[1] Spoken language systems not only make errors at segmental level, transcribing the consonants and vowels of the input, but they remain awkward to use and difficult to correct. Such systems are not yet fluent conversational partners.

My research lies at the intersection of computational linguistics, natural language processing, and spoken language processing. My work aims to elucidate and model human language use and processing and to incorporate these findings into systems for a wide range of applications, from spoken dialog systems to topic segmentation to information retrieval. My research aims to improve language understanding by exploiting linguistic information beyond the surface word or sentence level. The speech signal carries not only the segmental identity of phonemes, consonants and vowels, but also prosodic information, including pitch, loudness, and duration. Human languages make crucial use of prosody to convey information, for example, distinguishing a question from an answer with a rising intonation in English ("Yes?"). However, despite the fundamental importance of this prosodic information, computational approaches to speech recognition and processing have largely viewed such variation as a source of noise to be normalized away.

My work employs machine learning and computational models to capture acoustic evidence from pitch, intensity, and duration to improve automatic understanding of spoken language. This improvement comes from word-level information, such as through recognition of tone in Mandarin Chinese, and from information about the longer range context of extended spans of speech, through recognition of topical segments in broadcast news or miscommunication in human-computer dialog. These computational analyses can further provide new insight into the linguistic structure of tone and intonation.

Prosody and Miscommunication in Human-Computer Dialog

My work in human-computer dialog has demonstrated the importance and utility of prosodic cues to understanding in spoken language systems. Miscommunication in human-computer interaction

[1] http://www.computerworld.com/action/article.do?command=printArticleBasic&articleId=9012345

is unavoidable, and thus techniques to facilitate detection and correction of miscommunication are crucial. Consider the following example from the stock quotes application of the SpeechActs *[Yankelovich et al., 1995]* spoken dialog system:

1 User: Give me the price for AT&T.
2 System: Hewlett-Packard was 83 3/4,
up 2 1/2 since yesterday.
3 User: Give me the price for *AT&T*.

The system has clearly misrecognized the user's input in [1], substituting HP for AT&T. The user in [3] repeats their original input in an attempt to correct this misrecognition. We will refer to [3] as an instance of a "spoken correction" or a "repeat correction". While [1] and [3] have the same word sequence, [3] is spoken more slowly, with more pausing, and with greater emphasis on "AT&T." These sorts of interactions are both common in spoken language systems and very frustrating to the user. The system, on the other hand, is unaware that an error has occurred. Such errors are often difficult to correct, leading to extended sequences of misrecognitions and attempted corrections, termed "error spirals." Due to the constrained language understood by current spoken language systems, many of the lexical and syntactic cues generally exploited in human-human interaction to signal the presence of error or the need for a correction are unavailable or underutilized. In a field trial of the SpeechActs system, users cursed more often than they used an expressions such as "no, I meant" to formulate a correction. Since these textual cues are unavailable or misrecognized, we must go beyond the word level to solve this problem.

Extending acoustic-prosodic analysis of simulated errors in Wizard-of-Oz studies in *[Oviatt et al., 1996]*[2] and *[Oviatt et al., 1998b]*, my research demonstrated that rather than simply being a source of noise, the systematic changes in speaking style associated with corrective utterances to computers - often called hyperarticulation - can be used to automatically identify breakdowns in human-computer communication. The features can be exploited by machine learning classifiers for detection of spoken corrections. In [Levow, 1998], I employed duration, pause, and pitch features to train a decision tree classifier to identify spoken corrections at 77% accuracy on a balanced test set, approaching human performance on this task. Based on this work, comparable findings were subsequently identified by other researchers across a range of languages including German, Dutch, and Swedish and extended and integrated with recognizer confidence scores to further improve detection of corrections [Hirschberg et al., 2004]. Levow ([Levow, 1999, Levow, 2002]) extended this work to demonstrate that the acoustic-phonetic and durational adaptations present in spoken corrections represent divergences from more general models of conversational speech, suggesting the need for adaptive phonetic and durational models and partially explaining the increased probability of misrecognition after a misrecognition resulting in "error spirals".

As in the example above, many recognition errors are "local" or "focal", involving a substitution of a single item for another. In my recent research, building on studies in *[Oviatt et al., 1998a]* about local corrections, I have further found that the specific word of the utterance being corrected in a focal correction can be identified automatically. In particular, the largest pitch maximum, pitch range, and highest intensity (associated with perceived loudness) in the utterance are associated with the word being corrected. These features were identified and combined for classification within a boosting framework, yielding an accuracy of 85.5% in distinguishing locally corrected elements from other words, an almost 50% reduction in error from most common class assignment [Levow, 2004c].

[2]Citations in italics refer to papers on which I am a co-author, but not first author.

Context and Learning in Multilingual Tone and Pitch Accent Recognition

Prosodic cues convey a wealth of information in speech. In so-called "tone languages" or languages with "lexical tone", such as Mandarin Chinese and isiZulu (a Nguni language), the pitch of a syllable determines its meaning, in conjunction with segmental content, that is, the sequences of consonants and vowels. In Mandarin, for example, the syllable "ma" can mean "mother", "hemp", "horse", or "scold", depending on whether it is spoken with a high-level, rising, low, or high-falling pitch contour. In English, in contrast, a "pitch accent" on a syllable contributes to a word's prominence in an utterance and has been associated with emphasis or marking new or important information. Clearly, for speech recognition and language understanding, effective tone and pitch accent recognition are desirable; however, many speech recognition systems for Mandarin do not explicitly recognize tone.

My research aims to address two significant challenges in tone and pitch accent recognition, that have limited the accuracy and applicability of previous approaches. The first challenge is the effect of context on tone and pitch accent recognition. Tone and pitch accent realization are both relative to context such speaker pitch and also affected by local coarticulation at the pitch level. Tonal coarticulation occurs when the articulators, which control the rate of vocal fold vibration and hence speaking pitch, transition continuously, rather than instantaneously, from one pitch target to another, due to physical constraints. These effects can be extreme enough to change a falling tone to a rising tone, when preceded by a low target and followed by a high one. The second challenge is posed by the need for large quantities of training data to support data intensive supervised machine learning techniques, such as Support Vector Machines or Decision Trees. This manually labeled training data can be expensive and time-consuming to produce, relying on expert labelers. The need for such costly training data has limited the genres and styles of speech which have been explored. My research employs a common prosodic model and classification framework for lexical tone, in Mandarin Chinese and isiZulu, as well as English pitch accent, inspired by the Pitch Target Approximation model [Xu and Wang, 2001] to describe coarticulatory effects.

Within this framework using Support Vector Machine classifiers, this approach demonstrates tone and pitch accent accuracies competitive with those reported in the literature for prosody-only classification, of 76.5% for Mandarin, 76% for isiZulu, and 81.3% for English pitch accent [Levow, 2005a]. I find not only that any contextual modeling improves over classification without context features, but also that recognition based on preceding context yields greater significant improvements than that based on following context. These findings are consistent across all three languages, even though they represent radically different language families, suggesting a fundamental cross-lingual constraint. This capture of contextual evidence both improves automatic tone and pitch accent recognition, and also sheds light on phonetic theory, indicating a greater role of carryover coarticulation from a preceding syllable than anticipatory coarticulation with following syllables, consistent with [Xu, 1997]. Additional acoustic features, such as overall intensity and band energy, enhance coarticulatory modeling *[Wang and Levow, 2006]* and recognition of the extremely challenging, context-dependent Mandarin neutral tone.*[Surendran and Levow, 2006b, Surendran and Levow, 2006a]*

To address the challenge of training data demands, we exploit unsupervised and semi-supervised machine learning techniques. Unsupervised clustering requires no manually labeled training data to cluster samples. Using unsupervised clustering, such as standard k-means and several spectral

clustering techniques, the best approaches achieve 75% (Mandarin Chinese) to over 90% (English and isiZulu) of supervised accuracy, which required hundreds to thousands of manually labeled training examples. Similarly, semi-supervised machine learning techniques employ a small number of labeled examples in conjunction with a larger set of unlabeled examples to train the classifier, thereby reducing training data requirements. Experiments exploiting a semi-supervised approach using Laplacian SVMs [Belkin et al., 2004] achieved over 90% of fully supervised levels for pitch accent and tone recognition, using dramatically less training data than the supervised approaches. These results demonstrate that the underlying structure of the acoustic space for tone and pitch accent categories can be effectively exploited to reduce the need for costly labeled training data [Levow, 2006].

This research represents one of the first efforts in unsupervised and semi-supervised approaches to tone and pitch accent recognition. Analysis of these automatically derived clusters will provide insight into the relationship between categories suggested by the intrinsic acoustics and prosodic inventories proposed by linguistic theories. Reduced training data requirements will allow exploration of tone and intonation in a wider range of languages and speaking styles, and improvements in tone recognition accuracy can enhance speech recognition in tone languages.

Prosody in Discourse and Dialog Structure

Most linguistic and computational linguistic work has focused on the level of the word or sentence. However, sentences or utterances do not generally occur in isolation, but rather form coherent extended spans of text or speech, termed "discourse". "Dialog" concerns spoken interactions with two or more participants. Discourse is not uniform or monolithic, but rather structured, and is more than just the sum of its parts. In well-written discourse, structure is clearly signaled by chapters, sections, and paragraphs; turns in dialog are likewise separated. While the same structure is present in speech, the orthographic cues are absent; however, the structure can be identified not only from the words of the utterances, but also, crucially, from intonational cues. For example, one would expect a silence or pause between stories in a news broadcast.

Successful segmentation of a discourse facilitates correct interpretation of pronouns and other references, summarization and retrieval, and can constrain interpretation. For example, to support search and retrieval from a speech source such as CNN broadcast news, segmentation into individual stories would be a necessary first step; playback of an entire broadcast in response to search for a single term would be far too cumbersome. Likewise in dialog, it has been argued ([Duncan, 1974, Sacks et al., 1974]) that turn-taking is orderly; speakers do not interrupt each other at arbitrary points, but instead opportunities for turn exchanges are signaled by the speaker. Identification of these signals will enable spoken language systems to more fluently converse, rather than simply waiting for some arbitrarily defined silence interval to be reached.

The identification of discourse segments itself relies on a combination of lexical and prosodic cues. Swerts [Swerts, 1997] demonstrated that human annotators are able to more sharply and reliably identify segment boundaries in narrative when both text and audio were available than with text alone. Although some notable research has employed both text and prosodic cues for story segmentation of English broadcast news, the majority of approaches for broadcast news segmentation rely on text and possibly silence cues. While previous work focused on English monologue and human-human dialog, my research has instead considered Mandarin Chinese. Since Mandarin Chinese uses the pitch contour on syllables to determine the meaning of a word (as discussed

above), it was unclear whether prosodic cues such as pitch would be available in Mandarin to signal discourse and dialog structure, as they are in English. First, I demonstrated significant differences in pitch height and intensity for topic initial versus topic final ([Levow, 2004b, Levow, 2004d, Levow, 2004a]), and turn-unit initial versus turn-unit final positions [Levow, 2005b], with words and syllables in initial positions exhibiting significantly higher pitch and intensity than those in final position. Comparable experiments show that these contrasts are similar to those for English, although the effects of apparent loudness are greater in Chinese. Furthermore, these contrasts support automatic recognition of topic and turn boundaries in Mandarin Chinese, based on prosodic evidence alone. Furthermore, [Levow, 2004a] demonstrates that even without explicit silence cues, prosodic cues can more robustly and sharply identify segment boundaries than text cues alone, and the best overall effectiveness for automatic segmentation of Voice of America Mandarin broadcasts required majority agreement on boundary decisions from three feature-specific classifiers for each of text, silence, and other prosodic cues [Levow, 2004b].

The effectiveness of prosodic cues for topic and turn boundary recognition, however, raises questions about the interaction of tone and intonation since Mandarin is jointly employing pitch to indicate discourse initiation and finality as well as lexical tone. Contrastive analysis of tone in topic and turn-unit final and non-final positions demonstrates that pitch contours and relative heights that determine tonal identity are largely preserved, although all tones in final positions are lower in pitch than their corresponding non-final instances. Thus, given discourse structure status, tone is recognizable.

Improving Access to Information

My work in information retrieval has also exploited information beyond the surface text form of the document or specification of the information need to improve retrieval. In information retrieval, the goal is to connect the user's specification of an information need to documents that meet that need. A fundamental challenge is that straight-forward text matching may be stymied by differences in lexical choice as well as differences in language or medium of the document.

In work on cross-language and spoken document retrieval, I have employed English language queries to retrieve documents in a wide range of languages, in text and speech, including Mandarin Chinese, Japanese, French, German, and other European languages, in the context of several international information retrieval evaluations [Levow et al., 2005]. A key insight in this work has been that the original surface form may not optimally support matching and that there may not be a single surface form which does. I focus on the development of techniques that are as language-independent and portable as possible ([Levow et al., 2001], *[Resnik et al., 2001]*), requiring only simple linguistic resources. Using a dictionary-based term-for-term translation approach, rather than identifying a single optimal translation, we developed approaches to incorporate and suitably weight multiple translation alternatives [Levow and Oard, 2002]. Query and document "expansion" procedures identified terms from highly ranked, presumed relevant retrieved documents to enrich the representation both preceding and following translation, with topically related terms not in the original, to overcome gaps in translation resources and mitigate the effect of lexical choice ([Levow, 2003b, Levow, 2003a, Lo et al., 2003]). Finally, approaches employed mull-scale indexing and retrieval (*[Meng et al., 2001, Meng et al., 2004]*) to improve effectiveness by eschewing words as the primary unit of representation in favor of character sequences and phrasal units. By going beyond the original text content, these approaches support cross-lingual information retrieval that

can improve not only over baseline term-for-term translation, but also over monolingual retrieval with original source queries.

In recent research on text classification and segmentation, we have explored a hybrid indexing approach within a spectral embedding framework. First, we observe that words with different parts of speech, such as common nouns, proper nouns, verbs, etc., have different natural measures of similarity. For example, for person names, a term matching approach seems appropriate, whereas for common nouns a smoothed notion of concept similarity based on cooccurrence would be more suitable. Second, we employ a spectral embedding approach based on decomposition of a matrix of linguistically motivated pairwise term similarities for common nouns. Document representations are then a combination of the representations for the common nouns in the document and for all other terms. This hybrid indexing with spectral embedding can significantly improve over baseline vector space representations as well as Latent Semantic Analysis when used to compute similarities for text categorization ([Matveeva and Levow, 2007b] and segmentation ([Matveeva and Levow, 2007a]).

These approaches all improve over simple term matching by exploiting more evidence and creating richer representations than the surface forms alone allow.

Ongoing Research and Future Directions

Fundamentally, my research goal is to understand and model the role and process of prosody in human language and to employ these models to improve the capabilities and usability of spoken language understanding systems. This task requires connecting acoustics - from duration, pitch, and intensity - with high-level linguistic information, such as lexical tone, phrase or sentence boundaries, and discourse structure and function. This work brings together information beyond the segmental level and linguistic structure beyond the word or sentence. My research thus far has explored several of these connections, though relatively independently. My current and future work extends this investigation into prosody and language understanding, both integrating these threads and exploring new avenues.

Context, Tone, and Intonation

Context clearly plays an important role in the realization of tone and pitch accent. Modeling local coarticulatory and phrasal context improves recognition of these prosodic events. However, prosody in natural, fluent discourse and dialog depends not only on this local context, but also on the broader communicative context. Since prosodic features convey a wealth of information in spoken language, from word meaning to discourse structure, pitch, intensity, and duration must simultaneously encode different types of information. What is the role of intonation in a language with lexical tone, where pitch on syllables determines meaning? If questions and topic boundaries can affect pitch, how can tone still be understood by a native speaker? By an automatic classifier? If turn and topic boundaries coincide, can the discourse and dialog structure both be realized prosodically? How do words together with prosody resolve potential ambiguity? In a multimodal context, where gaze and gesture can also provide turn-taking and discourse information, what is the role of prosody and how is it affected?

Building on coarticulatory models and studies of production and perception, I will develop systems that can jointly model and recognize tone and intonational information at many levels. This work will integrate and extend disparate studies of both textual and prosodic evidence for phrase

and sentence boundaries, discourse structure, turn-taking, and dialog acts as well as tone and pitch accent recognition. Processes operating at different scales, from the phoneme to the syllable to the text span, will be incorporated. Analysis of spoken monologue, dialog, and multimodal discourse will highlight the interactions of prosody with different discourse contexts. This modeling will support both contextually appropriate synthesis and context-sensitive recognition, enabling spoken language systems that are more capable conversational partners.

Intonation and Language Learning

Intonation plays an important role in language learning. Very young infants are sensitive to the rhythm of different languages and prosodic cues to different word classes. Prosody has been proposed as a feature for many tasks from word learning to segmentation of the acoustic stream.

I will explore the role of intonation in language learning as well as the learning of intonation. How do children, apparently effortlessly, isolate the tone and intonational events in the acoustic stream? Does the structure of child-directed speech aid in the language learning process, and, if so, how? Can these cues be exploited to bootstrap computational language learning? Can second language learners acquire the same tone and intonational targets as first language learners?

Investigation in this area will draw on linguistic theory and psychological and developmental studies, in conjunction with computational models of intonation and machine learning. This research will develop classifiers and detectors to identify intonational events in the acoustic stream with little supervised training, as is believed to be the case in first language acquisition. The work will assess learnability from different speaking styles, including child-directed speech, and will explore whether the prosodic structure facilitates segmental learning and recognition. It will also compare the tone and intonation of second language learners with native speakers and support the development of tools for second language learners in tone and intonation.

With many deployed spoken language systems restricting the user to simple interactions, one could argue that there is little need for prosody. However, as we move beyond such constrained systems, the role of prosody becomes increasingly important. To achieve full conversational fluency and robustness with natural conversational input, systems must not only achieve high word recognition accuracy but also need to approach human levels of discourse and dialogue control and understanding and even become sensitive to indications of uncertainty or certainty, frustration, or anger. Only through understanding and integrating prosody will we be able to achieve these more natural conversational partners. My research aims to make such systems possible.

References

[Belkin et al., 2004] Belkin, M., Niyogi, P., and Sindhwani, V. (2004). Manifold regularization: a geometric framework for learning from examples. Technical Report TR-2004-06, University of Chicago Computer Science.

[Duncan, 1974] Duncan, S. (1974). *Some signals and rules for taking speaking turns in conversations*, pages 298–311.

[Hirschberg et al., 2004] Hirschberg, J., Litman, D., and Swerts, M. (2004). Prosodic and other cues to speech recognition failures. *Speech Communication*, 43(1-2):155–175.

[Levow, 1998] Levow, G.-A. (1998). Characterizing and recognizing spoken corrections in human-computer dialogue. In *Proceedings of COLING-ACL '98*, pages 736–742.

[Levow, 1999] Levow, G.-A. (1999). Understanding recognition failures in spoken corrections in human-computer dialogue. In *Proceedings of ESCA Workshop on Dialogue and Prosody*, pages 193–198.

[Levow, 2002] Levow, G.-A. (2002). Adaptations in spoken corrections: Implications for models of conversational speech. *Speech Communication*, 36(1–2):147–163.

[Levow, 2003a] Levow, G.-A. (2003a). Issues in pre- and post-translation document expansion: Untranslatable cognates and missegmented words. In *Proceedings of 4th International Workshop on Information Retrieval in Asian Languages*, pages 77–83.

[Levow, 2003b] Levow, G.-A. (2003b). Multi-scale document expansion for mandarin chinese. In *Proceedings of Workshop on Multilingual Spoken Document Retrieval, at International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2003*, pages 73–78.

[Levow, 2004a] Levow, G.-A. (2004a). Assessing prosodic and text features for segmentation of mandarin broadcast news. In *Proceedings of HLT-NAACL 2004 Workshop: Interdisciplinary Approaches to Speech Indexing and Retrieval*, pages 28–32.

[Levow, 2004b] Levow, G.-A. (2004b). Combining prosodic and text features for segmentation of mandarin broadcast news. In *Proceedings of the 3rd SIGHAN Workshop on Chinese Language Processing*.

[Levow, 2004c] Levow, G.-A. (2004c). Identifying local corrections in human-computer dialogue. In *Proceedings of International Conference on Spoken Language Processing 2004*.

[Levow, 2004d] Levow, G.-A. (2004d). Prosody-based topic segmentation for mandarin broadcast news. In *Proceedings of HLT-NAACL 2004, Companion Volume*, pages 137–140.

[Levow, 2005a] Levow, G.-A. (2005a). Context in multi-lingual tone and pitch accent prediction. In *Proceedings of Interspeech 2005*, pages 1809–1812.

[Levow, 2005b] Levow, G.-A. (2005b). Turn-taking in mandarin dialogue: Interactions of tone and intonation. In *Proceedings of the Fourth SIGHAN Chinese Language Processing Workshop*, pages 72–78.

[Levow, 2006] Levow, G.-A. (2006). Unsupervised and semi-supervised learning of tone and pitch accent. In *Proceedings of HLT-NAACL 2006*, pages 224–231.

[Levow and Oard, 2002] Levow, G.-A. and Oard, D. W. (2002). *Signal Boosting for Translingual Topic Tracking: Document Expansion and N-best Translation*, pages 175–196. Kluwer.

[Levow et al., 2001] Levow, G.-A., Oard, D. W., and Resnik, P. (2001). Rapidly retargetable interactive translingual retrieval. In *Proceedings of Human Language Technology Conference (HLT) 2001*, pages 294–298.

[Levow et al., 2005] Levow, G.-A., Oard, D. W., and Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management: Special Issue on Cross-language Information Retrieval*, 41(4).

[Lo et al., 2003] Lo, W.-K., Li, Y.-C., Levow, G.-A., Meng, H., and min Wang, H. (2003). Multi-scale document expansion in english-mandarin cross-language spoken document retrieval. In *Proceedings of the Conference of the International Speech Communication Association (Interspeech) 2003*, pages 2337–2340.

[Matveeva and Levow, 2007a] Matveeva, I. and Levow, G.-A. (2007a). Hybrid document indexing with spectral embedding. In *Proceedings of HLT-NAACL 2007*, pages 113–116.

[Matveeva and Levow, 2007b] Matveeva, I. and Levow, G.-A. (2007b). Topic segmentation with hybrid document indexing. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP) 2007*.

[Meng et al., 2001] Meng, H., Chen, B., Grams, E., Khudanpur, S., Levow, G.-A., Lo, W.-K., Oard, D. W., Schone, P., min Wang, H., and Wang, J. (2001). Mandarin-English (MEI): Investigating translingual speech retrieval. In *Proceedings of Human Language Technology Conference (HLT) 2001*, pages 239–245.

[Meng et al., 2004] Meng, H. M., Chen, B., Khudanpur, S., Levow, G.-A., Lo, W.-K., Oard, D. W., Schone, P., Tang, K., min Wang, H., and Wang, J. (2004). Mandarin-English Information (MEI): Investigating translingual speech retrieval. *Computer Speech and Language*, 18(2):163–179.

[Oviatt et al., 1996] Oviatt, S., Levow, G.-A., MacEachern, M., and Kuhn, K. (1996). Modeling hyperarticulate speech during human-computer error resolution. In *Proceedings of International Conference on Spoken Language Processing '96*, pages 801–804.

[Oviatt et al., 1998a] Oviatt, S., Levow, G.-A., MacEachern, M., and Moreton, E. (1998a). Modeling global and focal hyperarticulation during human-computer error resolution. *Journal of the Acoustical Society of America*, 104(5):1–19.

[Oviatt et al., 1998b] Oviatt, S., MacEachern, M., and Levow, G.-A. (1998b). Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication*, 24(2):87–110.

[Resnik et al., 2001] Resnik, P., Oard, D. W., and Levow, G.-A. (2001). Improved cross-language retrieval using backoff translation. In *Proceedings of Human Language Technology Conference (HLT) 2001*, pages 153–155.

[Sacks et al., 1974] Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735.

[Surendran and Levow, 2006a] Surendran, D. and Levow, G.-A. (2006a). Additional cues for mandarin tone recognition. Technical Report TR-2006-04, University of Chicago, Computer Science.

[Surendran and Levow, 2006b] Surendran, D. and Levow, G.-A. (2006b). Local rhyme-based acoustic features for mandarin tone recognition. Technical Report TR-2006-05, University of Chicago, Computer Science.

[Swerts, 1997] Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, 101(1):514–521.

[Wang and Levow, 2006] Wang, S. and Levow, G.-A. (2006). Improving tone recognition with combined frequency and amplitude modelling. In *Proceedings of Interspeech 2006*, pages 2386–2389.

[Xu, 1997] Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, 25:62–83.

[Xu and Wang, 2001] Xu, Y. and Wang, Q. E. (2001). Pitch targets and their realization: Evidence from mandarin chinese. *Speech Communication*, 33:319–337.

[Yankelovich et al., 1995] Yankelovich, N., Levow, G.-A., and Marx, M. (1995). Designing speechacts: Issues in speech user interfaces. In *Proceedings of CHI '95*, pages 369–376.