

GENERATING PITCH ACCENT DISTRIBUTIONS THAT SHOW INDIVIDUAL AND STYLISTIC DIFFERENCES

Janet E. Cahn

Massachusetts Institute of Technology
cahn@media.mit.edu

ABSTRACT

I describe a limited-resource approach to generating prosody that mediates text-based information through a model of attention and working memory, whose simulation parameters are quantitative. The main parameter quantifies recall. Varying it varies what counts as given and new in a text, and therefore, the pitch accents with which the text is uttered. Currently, the system produces prosody in three different styles of read speech – child-like, adult expressive, and knowledgeable – and individual variation within each. A comparison with natural data shows clear and predictable stylistic similarities, although not at significance. However, informal feedback is more forgiving, indicating that the prosody is both natural and expressive for consecutive phrases, but that work is still needed to make this effect consistent throughout the text.

1. INTRODUCTION

There is longstanding consensus that text-to-speech prosody could be better, that is, more natural, more interesting and more appropriate to the structure and semantics of the text. There is a growing awareness that more than one type of natural prosody is desirable as well. Recent work has begun to reflect this awareness, mainly for speaking style. For example, Johnson[13] varies lookahead to generate the prosody of read or spontaneous speech; Abe[1] isolates global and style-dependent rules to re-synthesize speech in the advertisement, novel (fiction) and encyclopedia (non-fiction) styles for Japanese. Other work ([16, 4, 6, 12]) addresses one source of variation within and among individual speakers, namely their emotional state or affective disposition.

The approach I describe generates three speaking styles and, in addition, individual differences within each style. With the exception of Johnson’s work, most other work tends towards a description rather than a production model. In my approach, the productive cause is the ability of a speaker to recall previously uttered items. The results of search and storage are mapped to intonation and timing. Varying the search parameter influences style; varying the storage parameter affects individual variation within a style. Currently, the model produces three styles likely to be associated with attentional and memory differences: a child-like exaggerated prosody for limited recall; a more adult but still expressive style for mid-range capacities;

and a knowledgeable style for greatest recall.

2. A MEMORY MODEL FOR PROSODY

Prosody organizes spoken text into phrases, and highlights its most salient components with *pitch accents*, distinctive pitch contours applied to the word. Pitch accents are both attentional and propositional. Their very use indicates salience; their particular form conveys a proposition about the words they mark. For example, speakers typically mark salient information with a high pitch accent (denoted as H*) if they believe it to be *new* to the addressee. Conversely, when they believe that the addressee is already aware of the information, they will typically de-accent it[3] or, if it is salient, apply a low pitch accent (L*)[19]. Re-stated as a commentary on working memory, the H* accent conveys the speaker’s belief that the addressee can not retrieve the accented information from working memory. De-accenting conveys the opposite expectation implicitly; the L* accent does so explicitly. This view predicts different speaking styles as a consequence of the speaker’s beliefs about an addressee’s storage and retrieval capacities. For example, it ascribes the exaggerated intonation that adults use with infants and young children[9] to the belief that the child’s knowledge and attention are extremely limited; therefore, he needs clear and explicit prosodic instructions as to how to process language and interaction.

The model of working memory I use shows how retrieval limits can determine the information status of an item as either given or new, and therefore, its corresponding prosody. It was developed and implemented by Thomas Landauer[15] and models *working memory* as a periodic three dimensional Cartesian space, the *focus of attention* via a moving search and storage *pointer* that traverses the space in a slow random walk, and *retrieval ability* via a *search radius* that defines the size of a region whose center is the pointer’s current location. Search for familiar items proceeds outward from the pointer, one city block per time step, up to the distance specified by the search radius.

As a consequence of the pointer’s slow random walk, incoming stimuli are stored in a spatial pattern that is locally random but globally coherent. That is, temporal proximity in the stimuli begets spatial proximity in the model. It contrasts with stack models of memory that are strictly

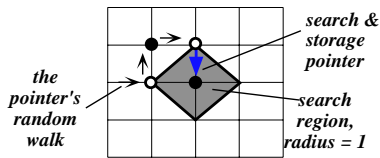


Figure 1: Using AWM, stimuli are classified as given if they have counterparts within the the search radius. New items have no such counterparts because they are either not in working memory, or are stored outside the radius.

chronological, and semantic spaces in which distance is conceptual rather than temporal. Most importantly, it is a valid computational model of attention and working memory (AWM, from here on). Landauer used it to reproduce the well-known learning phenomena of recency and frequency, in which subjects tend to recall stimuli encountered most recently or most frequently[15]. It has since been used by Walker[21] to show that resource-bound dialog partners will make a proposition explicit when it is not retrievable or inferable, despite having been previously mentioned.

Retrieval in AWM is the process of *matching* the current stimulus to the contents of the region centered around the pointer. The search radius determines the size of this region and therefore is the main AWM simulation parameter. If a match is found within the search region, the stimulus is classified as *given*, otherwise, it is *new*. Figure 1 illustrates this with the simple example of filled and unfilled circles, a 4x4 AWM space, and a search radius of one. At the center of the search region is the current stimulus, a filled circle. Because the region contains no other filled circles, it is classified as new. Had the stimulus instead been an unfilled circle, it would have been classified as given because a match is retrievable within the search radius. Or, alternatively, had the search radius been two instead of one, a matching filled circle would have been found, and the stimulus also classified as given.

The update rule for AWM is simple: STORE and SEARCH. In the storage step, the pointer moves to the next randomly-determined location and stores the current stimulus. A search for a matching item then proceeds outward from the pointer. I adopt Landauer’s proposal that the search proceeds in constant time. All items at the same distance from the pointer are effectively compared in parallel, and time increments once for each city block distance covered by the search. The search stops if a match is found; if no matches are found, the search extends to the edge of the search region. In LOQ, the implemented system that maps AWM activity to pitch and timing, the total time devoted to storage and search becomes the duration of the word. This mapping finds support in psychological studies that show decreased duration for more accessible words[10] and increased duration for rarer ones[11]).

2.1. Discussion

The ability to identify given and new items makes AWM a useful producer of prosody based on this distinction. Ostensibly, it shows how a speaker’s processing affects her

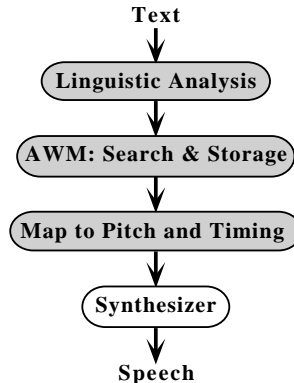


Figure 2: AWM as a component of the LOQ system.

prosody. However, although the working memory belongs to the speaker, its operation and determinations may reflect the speaker’s own retrieval capacities, her estimate of those of the addressee, or a mixture of both. That is, a speaker can always adapt her style (prosodic and lexical) to the needs of a less knowledgeable or capable addressee. A cooperative and communicative speaker will usually do this. However, she cannot model a retrieval capacity greater than her own – her own knowledge and attentional limits always constitute the upper bounds on her performance.

3. SYSTEM DESIGN

The AWM component is embedded in a software implementation, LOQ, that takes a text-to-speech approach. As shown in Figure 2, the input to AWM is text, the output is speech. Therefore, LOQ models read rather than spontaneous speech. Text comprehension is the process of searching for a match. Uttering the text is a question of mapping the search process and its results to prosodic features and sending the prosodically annotated text to the synthesizer.

As with commercial text-to-speech synthesizers, the text is analyzed before prosody is assigned. However, the LOQ analysis is richer. It takes advantage of on-line linguistic databases to approximate the speaker’s knowledge of English semantics, pronunciation and usage. The structural analysis is richer as well, providing both grammatical structure (subject, verb, object), empty categories (ellipses, for example) and information about clausal attachment. However, the main difference is that LOQ interposes a model of limited attention and working memory between the text analysis and prosodic mapping components.

3.1. Matching

For the example in Figure 1, the matching criterion is binary and simple – a circle is either filled or unfilled. However, language is many more times complex, and matches may occur for a variety of features, some of which are more informative than others. The matching criteria used in LOQ attempt to distill from the literature (e.g., [17, 8]) the most relevant and prevalent ways that items in mem-

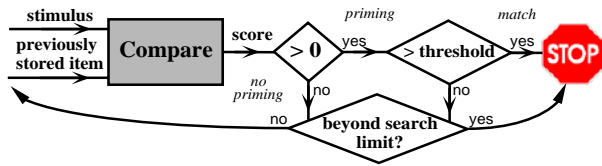


Figure 3: LOQ matching algorithm.

ory *prime* for the current stimulus, and by the same token, the ways in which the current stimulus can function as a *retrieval cue*. In other words, they gauge the mutual information between the current stimulus and previously stored items.

Altogether, LOQ tests for matches on twenty-four semantic, syntactic, collocation, grammatical and acoustical features. Each test contributes to the total match score, which is then compared to a *threshold*. If it is below, the search continues; if above, it stops. As shown in Figure 3, matches on any criterion express priming, and scores above the threshold constitute a match sufficient to stop the search even before it reaches the edge of the search region. Because some tests are more informative than others, a high score can reflect the positive outcome of many un-informative tests, or of one that is definitive. Thus, in the current ordering, co-reference ensures a match, while structural parallelism in and of itself does not.

3.2. The text input

The matching criteria determine the form and kind of information in the text input. As with commercial synthesizers, this includes part of speech tagging. LOQ uses the output of Lingsoft’s ENGCG (English Constraint Grammar) software which provides both tags and phrase structure information. However, reliable automatic means for identifying other information, such as grammatical clauses, empty categories, attachment and co-reference do not yet exist. Therefore, this information was entered by hand.

The LOQ software turns the parsed and annotated text into a sequence of tokens that assembles clauses in a bottom up fashion, starting with the word and followed by the syntactic and grammatical clauses to which it belongs. This models the reader’s assembly of the words into meaningful syntactic and grammatical groupings.¹

To facilitate the matching process, the text is also augmented with information from the WordNet database for semantic comparisons, a pronunciation database for acoustical comparisons and the Thorndike-Lorge and Kucera-Francis for word frequency counts² to scale the match score by the prior probability for the language. The WordNet synonym indices were assigned by hand. However, all subsequent semantic comparisons using WordNet are automatic as required by the matching process.

¹ Adapting this for a spontaneous speaker would proceed in reverse, from the concept, to grammatical roles, syntactic phrases and finally, the words.

² As provided in the Oxford Psycholinguistic Database.

3.3. Mapping to accent type

I have described how AWM can produce the L* accent (or none) for retrievable items, and H* for new ones. However, there are more than two pitch accents – Pierrehumbert *et al.*[2] identify six³ – and more components to prosody. Obtaining them from one model first requires an adjustment such that given or new status is determined from the effect of the stimulus on the region as a whole, as follows: The result of any one comparison affects the “state” of the item to which the stimulus is compared. State is simply defined – a L annotation records a match on most any criterion,⁴ and a H annotation records a match score of zero. Thus, the comparison process registers both priming and a true match. Both receive L annotations, but only a match whose score exceeds the threshold stops the search.

A pitch accent is then derived by comparing the contents of the search radius *before* and *after* the matching process. Majority rules apply such that the annotation with the higher count becomes the defining tone. If both the before and after configurations are composed mainly of L annotations, the accent form is L+L, which becomes the L* accent. However, if there is a change, for example, from a L to H majority, the accent form is L+H. The interpretation of L+L is, roughly, that a familiar item was expected and provided. Likewise, the interpretation for L+H is that a familiar item was expected but an unexpected one provided.

To complete the bitonal derivation, LOQ treats the location of the main tone as a categorical reflection of the magnitude of the effect of the stimulus. If the stimulus changes the annotations for the majority of items in the search region, the second tone is the main tone. Otherwise, it is the first. This schema produces the six pitch accents identified by Pierrehumbert *et al.* Operationally, it provides the model with a simple form of feedback – the results of prior processing persist and contribute to a bias that affects future processing.

The pitch accent mapping illustrates the main features of the prosodic mapping in general. First, all mappings reflect the activity and state within the region defined by the search radius. Second, they express some aspects of prosody as a plausible consequence of search and storage. For example, storage and search times are mapped to word and pause duration. However, others – for example, the bitonal derivation – are, at best, coherent with the operation and purpose of the model and not contradicted by the current (sparse) data on the relation of cognitive capacity to the prosody of read speech. In all, the mapping from AWM activity and state produces intonational categories (pitch accent, phrase accent and boundary tone) and their prominence, word duration, pause duration and the pitch range of an intonational phrase.

³ L*, H*, L+H*, L*+H, H+L*, H*+L.

⁴ Some criteria are parasitic and only contribute to the score in combination with other criteria.

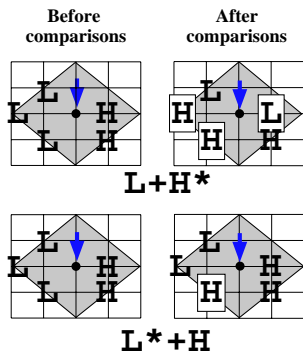


Figure 4: In LOQ, bitonals occur when the L/H counts differ before and after the matching process. The main tone of a bitonal is treated as a categorical indicator of the magnitude of the effect of the stimulus on the context.

4. SIMULATIONS

Simulations were run using text from three different genres (fiction, news broadcast, rhymed poetry). The rhymed poetry produced the most phrases, thus showing the influence of genre. However, across texts, the prosody was mainly influenced by the simulation parameters. For this reason, and because human comparison data are available for the news story (from the BU corpus[18]), results are reported only for this text.

The three simulation parameters are *search radius*, *pointer step size* and *memory size*. The search radius affects recall, and therefore has the strongest influence, especially on pitch accent type. Increasing the step size, which affects the sparseness of the distribution of items in memory, expands the range of search radii for which variation is greatest. Memory size, which affects global sparseness, appears to have little effect for the range of sizes I tested (from ten to fifty). Dimensionality – two, three and four – also appear to have little effect. Therefore, I report the results of two dimensional simulations, which run most quickly, and for the 50x50 memory, which most completely shows the behavior of the model as the search radius and step size are increased.

Five simulations were run per parameter configuration. The search radii ranged from one to fifty and the step sizes from one to three. This produces 150 simulation groups in all, effectively, 150 “speakers”. The total number of simulations is 750 (5 runs x 3 step sizes x 50 radii). Differences within a simulation group model *intra-speaker variation*; differences across groups model *inter-speaker variation*. The text for the simulations is the first paragraph of the news story, comprised of sixty-eight words and four sentences.

5. RESULTS

I report the results only for the pitch accents, which are the canonical exemplar for the development of this ap-

proach. The pitch accent distributions⁵ shown in Figure 5 bear out the limited-capacity predictions. As the search radius increases, the mean⁶ number of unaccented words increases as well, showing that more items become given as recall increases. Because the text is news, and the information new to the hearer, the H* accent is likely to be the most prominent accent. This is borne out in the simulations and in addition, is also found in the distributions in the natural data, as shown in Figure 6.

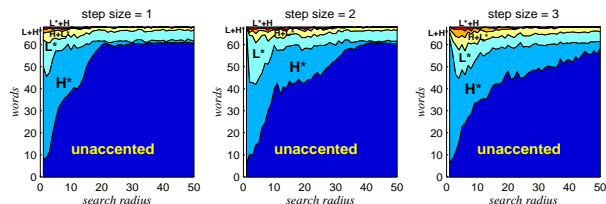


Figure 5: Mean distributions for accenting phenomena as a function of search radius and step size.

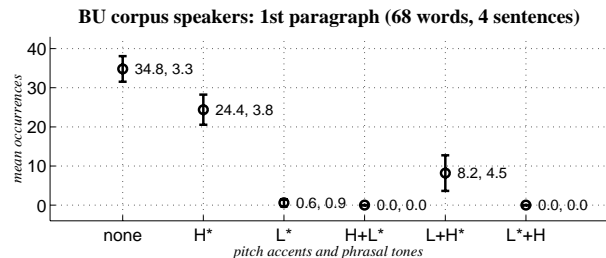


Figure 6: Mean and standard deviation among the accent type distributions for the natural speakers.

Accenting is also a function of step size. As shown by Figure 5, there are more accents for the larger step sizes. This is a consequence of the more sparse distribution (fewer items are retrievable within the radius) which weakens the link between temporal proximity in the input stream and spatial proximity in AWM.

To get a better sense of the variation within each simulation group, I used the kappa statistic[7, 5]⁷ to measure the agreement on (i) accent location and (ii) accent type, where six are considered – unaccented, H*, L*, L+H*, L*+H, H+L* (downstepped accents are included with non-downstepped forms).

A kappa of 1 indicates perfect agreement and therefore no variation among the speakers. A kappa of 0 indicates agreement at chance, and negative kappas indicate disagreement greater than chance. Krippendorff[14] maintains that a kappa above .8 indicates significant agreement, while kappa between .67 and .8 indicates that only tentative conclusions may be drawn. Kappa is typically used to determine reliability. However, I use it as an indicator of variability. Kappa values below Krippendorff’s criteria

⁵Using the five ToBI accent categories - -H*+L becomes a H* pitch and L phrase accent.

⁶The standard deviations are small, hence the means are meaningful.

⁷The kappa statistic measures agreement on categorical judgments among different coders.

demonstrate insignificant agreement, and therefore, variability.

For each simulation group, the kappa values were calculated first for pairwise comparisons and then averaged. The mean and standard deviations for each test are shown in Figure 7. They too show the influence of search radius and pointer step size. A radius of one scores high, mainly because the chance of variation is minimal.⁸ The radii between two and ten show the most within-speaker variation – their kappas for accent location and type are low.

Increasing the step size expands the range in which kappa is not significant and within-speaker variation greatest. As Figure 7 shows, for both tests and a step size of one, it becomes significant at a radius of eighteen; for a step size of two, at a radius of thirty-five; for a step size of three, at fifty.

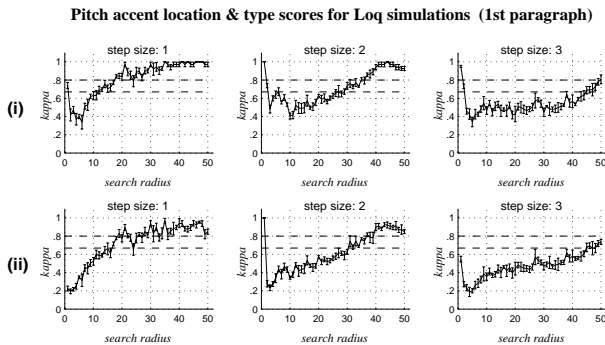


Figure 7: Mean kappa and standard deviations for LOQ simulations.

6. EVALUATION

The naturalness of synthetic prosody is difficult to evaluate via in perceptual tests[20]. However, informal comments from listeners revealed that while the three styles were recognizable and the prosody more natural-sounding than the commercial default, it was best over shorter sections rather than for the passage as a whole. The kappas for the comparison with the natural prosody (Figure 8) show that no score is significant or even tentatively so. However, the statistics for comparisons between the natural speakers are not much better. They are shown in Table 1.

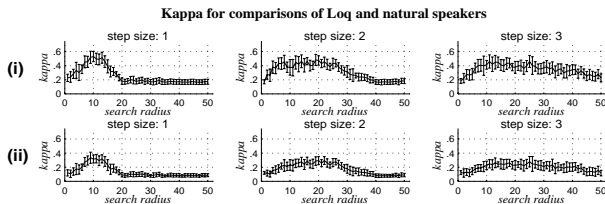


Figure 8: Mean kappa and standard deviations for the comparison between LOQ and natural speakers.

⁸However, the perfect score for the radius of two is an artifact of the interaction of the mapping with the Cartesian grid and is not meaningful.

Test	mean	mean – standard deviation
(i) Location	.57	.41
(ii) Type	.64	.48

Table 1: Kappas for comparisons between the natural speakers.

Nonetheless, the pattern for the highest scores is encouraging. Clearest for the step size of one, it shows the highest scores between radii of seven and fourteen. This accords with the reasonable expectation that the intonation of the newscasters would most resemble the knowledgeable style in LOQ (a search radius around ten), but, because the intent is to communicate with listeners unfamiliar with the material, a prosody that is positioned close to the adult expressive style. Looking at the scores for individual comparisons shows that some scores exceed the minimum thresholds. Their counts are shown in Table 2. The most comparisons succeed for a radius of nine, a step size of one and for two of the five speakers (designated M2B and F3A in the corpus).

Criterion	(i) Location	(ii) Type
mean – standard deviation	79	16
mean	16	1
> .67	4	0

Table 2: Number of kappa scores for individual comparisons between the LOQ and natural speakers that fall within and above the kappas for the natural data.

7. SUMMARY

The LOQ simulations show a mapping mainly from recall ability to pitch accent type. This produces stylistic variation in the output. As the consequence of the stochastic storage algorithm, within-speaker variation is also produced. A comparison with the natural data shows that pitch accent location and type distributions do not agree at or near significance. However, the comparison scores are not much worse than those within the natural data. Moreover, there are clear patterns for the recall values that produce the best matches to the natural data. This suggests that the continued pursuit of a limited-capacity approach to speech synthesis is worthwhile.

8. REFERENCES

1. Masanobu Abe. Speaking Styles: Statistical Analysis and Synthesis by a Text-to-Speech System. In Jan P. H. van Santen and Richard W. Sproat and Joseph P. Olive and Julia Hirschberg, editor, *Progress in Speech Synthesis*, chapter 39, pages 495–510. Springer-Verlag, 1996.
2. Mary E. Beckman and Janet B. Pierrehumbert. International structure in Japanese and English. In Colin Ewen and John Anderson, editors, *Phonology Yearbook 3*, pages 255–309. Cambridge University Press, 1986.
3. Gillian Brown. Prosodic Structure and the Given/New Distinction. In A. Cutler and D. R. Ladd, editors, *Prosody: Models and Measurements*, chapter 6, pages 67–77. Springer Verlag, 1983.

4. Janet E. Cahn. The Generation of Affect in Synthesized Speech. *Journal of the American Voice I/O Society*, 8:1–19, July 1990.
5. Jean Carletta. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254, 1996.
6. R. Carlson, B. Granström, and L. Nord. Experiments with emotive speech - Acted utterances and synthesized replicas. In *Proceedings*, pages 671–674, Banff, Alberta, Canada, October 1992. Second International Conference on Spoken Language Processing.
7. Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, XX(1):37–46, 1960.
8. David Fay and Anne Cutler. Malapropisms and the Structure of the Mental Lexicon. *Linguistic Inquiry*, 8(3):505–520, Summer 1977.
9. A. Fernald and T. Simon. Expanded Intonation Contours in Mother's Speech to Newborns. *Developmental Psychology*, 20:104–113, 1984.
10. Carol A. Fowler, Elena T. Levy, and Julie M. Brown. Reductions of Spoken Words in Certain Discourse Contexts. *Journal of Memory and Language*, 37:24–40, 1997.
11. Gina Geffen and Mary A. Luszcz. Are the spoken durations of rare words longer than those of common words? *Memory & Cognition*, 11(1):13–15, 1983.
12. Norio Higuchi, Toshio Hirai, and Yoshinori Sagisaka. Effect of Speaking Style on Parameters of Fundamental Frequency Contour. In Jan P. H. van Santen and Richard W. Sproat and Joseph P. Olive and Julia Hirschberg, editor, *Progress in Speech Synthesis*, chapter 33, pages 417–428. Springer-Verlag, 1996.
13. M. E. Johnson. Synthesis of English Intonation Using Explicit Models of Reading and Spontaneous Speech. In *Proceedings*. International Conference on Spoken Language Processing, 1996.
14. Klaus Krippendorff. *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, California, 1980.
15. Thomas K. Landauer. Memory Without Organization: Properties of a Model with Random Storage and Undirected Retrieval. *Cognitive Psychology*, 7:495–531, 1975.
16. I. R. Murray, J. L. Arnott, and A. F. Newell. Hamlet - simulating emotion in synthetic speech. In *Speech '88; Proceedings of the 7th FASE Symposium*. Institute of Acoustics, Edinburgh, 1988.
17. S. G. Nooteboom and J. M. B. Terken. What Makes Speakers Omit Pitch Accents? An Experiment. *Phonetica*, 39:317–336, 1982.
18. M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel. The Boston University Radio News Corpus. Technical report, Boston University, February 1995.
19. Janet B. Pierrehumbert and Julia Hirschberg. The Meaning of Intonation Contours in the Interpretation of Discourse. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, pages 271–311. MIT Press, 1990.
20. K. Ross and M. Ostendorf. Prediction of abstract prosodic labels for speech synthesis. *Computer Speech and Language*, 10:155–185, 1996.
21. Marilyn A. Walker. Testing collaborative strategies by computational simulations: cognitive and task effects. *Knowledge-Based Systems*, 8(2-3):105–116, April-June 1995.

GENERATING PITCH
ACCENT DISTRIBUTIONS THAT SHOW INDIVIDUAL
AND STYLISTIC DIFFERENCES

Janet E. Cahn

Massachusetts Institute of Technology
cahn@media.mit.edu

I describe a limited-resource approach to generating prosody that mediates text-based information through a model of attention and working memory, whose simulation parameters are quantitative. The main parameter quantifies recall. Varying it varies what counts as given and new in a text, and therefore, the pitch accents with which the text is uttered. Currently, the system produces prosody in three different styles of read speech – child-like, adult expressive, and knowledgeable – and individual variation within each. A comparison with natural data shows clear and predictable stylistic similarities, although not at significance. However, informal feedback is more forgiving, indicating that the prosody is both natural and expressive for consecutive phrases, but that work is still needed to make this effect consistent throughout the text.