ENCYCLOPEDIA OF LANGUAGE AND LINGUISTICS SECOND EDITION
CONTRIBUTORS' INSTRUCTIONS

PROOFREADING

The text content for your contribution is in final form when you receive proofs. Read proofs for accuracy and clarity, as well as for typographical errors, but please DO NOT REWRITE.

At the end of your article you will find a page which will contain several non-print items: abstract, author biographies and photographs, captions for any multimedia components, keywords (for indexing purposes), and the full contact details of each author. Full addresses are used to keep our records up-to-date (they will not appear in the published work) – for the lead author, this is the address that the honorarium will be sent. Please ensure that all of these items are checked thoroughly.

Titles and headings should be checked carefully for spelling and capitalization. Please be sure that the correct typeface and size have been used to indicate the proper level of heading. Review numbered items for proper order – e.g., tables, figures, footnotes, and lists. Proofread the captions and credit lines of illustrations and tables. Ensure that any material requiring permissions has the required credit line.

Any copy-editor questions are presented in an accompanying Manuscript Query list at the end of the proofs. Please address these questions as necessary. While it is appreciated that some articles will require updating/revising, please try to keep any alterations to a minimum. Excessive alterations may be charged to the contributors.

Note that these proofs may not resemble the image quality of the final printed version of the work, and are for content checking only. Artwork will have been redrawn/relabelled as necessary, and is represented at the final size.

PLEASE KEEP A COPY OF ANY CORRECTIONS YOU MAKE.

DESPATCH OF CORRECTIONS

Proof corrections should be returned in one communication to your academic editor JENNIFER LAI by 22-Apr-2005 using one of the following methods:

1. If corrections are minor they should be listed in an e-mail to lai@watson.ibm.com A copy should also be sent to: lali_proofs@elsevier.com. The e-mail should state the article code number in the subject line. Corrections should be consecutively numbered and should state the paragraph number, line number within that paragraph, and the correction.

2. If corrections are substantial, send the amended hardcopy by courier to JENNIFER LAI, IBM Research, 19 Skyline Drive, Hawthorne, NY 10532, USA, with a copy by fax to the Elsevier MRW Production Department (fax number: +44 (0)1865 843974). If it is not possible to courier your corrections, fax the relevant marked pages to the Elsevier MRW Production Department with a covering note clearly stating the article code number and title.

Note that a delay in the return of proofs could mean a delay in publication. Should we not receive corrected proofs within 7 days, the editors and Elsevier will proceed without your corrections.

CHECKLIST

Manuscript queries addressed/answered? ❑

Affiliations, names and addresses checked and verified? ❑

'Further Reading' section checked and completed? ❑

Permissions details checked and completed? ❑

Outstanding permissions letters attached/enclosed? ❑

Figures and tables checked? ❑

If you have any questions regarding these proofs please contact the Elsevier MRW Production Department at: lali_proofs@elsevier.com.

Article Number: LALI: 00914

a0005

Speech Synthesis, Prosody

J Hirschberg, Columbia University, New York, NY, USA

p0005 Text-to-speech (TTS) systems take unrestricted text as input and produce a synthetic spoken version of that text as output. During this process, the text input must be analyzed to determine the prosodic features that will be associated with the words that are produced. For example, if a sentence of English ends in a question mark and does not begin with a WH-word, that sentence may be identified as a yes–no question and produce with a rising 'question' contour. If the same word occurs several times in a paragraph, the system may decide to realize that word with less prosodic prominence on these subsequent mentions.

p0010 These decisions are known as 'prosodic assignment decisions.' Once they have been made, they are passed along to the prosody modeling component of the system to be realized in the spoken utterance by specifying the appropriate pitch contour, amplitude, segment durations, and so on. Prosodic variation will differ according to the language being synthesized and also according to the degree to which the system attempts to emulate human performance and succeeds in this attempt.

s0005 Issues for Prosodic Assignment in TTS

p0015 No existing TTS system, for any language, controls prosodic assignment or its realization entirely successfully. For most English synthesizers, long sentences that lack commas are uttered without 'taking a breath' so that it is almost impossible to remember the beginning of the sentence by the end; synthesizers that do attempt more sophisticated approaches to prosodic phrasing often make mistakes (e.g., systems that break sentences between conjuncts overgeneralize to phrasing such as 'the nuts | and bolts approach'). Current approaches to assigning prosodic prominence in TTS systems for pitch accent languages, such as English, typically fail to make words prominent or nonprominent as human speakers do. Since many semantic and pragmatic factors may contribute to human accenting decisions and these are not well understood, and since TTS systems must infer the semantic and pragmatic aspects of their input from text alone, attempts to model human performance in prominence decision have been less successful than modeling phrasing decisions. For most systems, the basic pitch contour of a sentence is varied only by reference to its final punctuation; sentences ending with '.', for example, are always produced with the same standard 'declarative' contour, contributing to a numbing sense of monotony. Beyond these sentence-level prosodic decisions, few TTS systems attempt to vary such other features as pitch range, speaking rate, amplitude, and voice quality in order to convey the variation in intonational meaning that humans are capable of producing and understanding.

p0020 Many TTS systems have addressed these issues using more or less sophisticated algorithms to vary prominence based on a word's information status or to introduce additional phrase boundaries based on simple syntactic or positional information. Although some of these algorithms have been constructed 'by hand,' most have been trained on fairly large speech corpora, hand-labeled for prosodic features. However, since prosodic labeling is very labor-intensive, and since the variety of prosodic behavior that humans vary in normal communication is very large and the relationship between such behaviors and automatically detectable features of a text is not well understood, success in automatic prosodic assignment in TTS systems has not improved markedly in recent years. Failures of prosodic assignment represent the largest source of 'naturalness' deficiencies in TTS systems today.

Prosody in TTS Systems

s0010

p0025 Prosodic variation in human speech can be described in terms of the pitch contours people employ, the items within those contours that people make intonationally prominent, and the location and importance of prosodic phrase boundaries that bound contours. In addition, human speakers vary pitch range, intensity or loudness, and timing (speaking rate and the location and duration of pauses) *inter alia* to convey differences in meaning. TTS systems ideally should vary all these dimensions just as humans do.

p0030 To determine a model of prosody for a TTS system in any given language, one must first determine the prosodic inventory of the language to be modeled and which aspects of that inventory can be varied by speakers to convey differences in meaning: What are the meaningful prosodic contrasts in this language? How are they realized? Do they appear to be related (predictable) in some way from an input text? How does the realization of prosodic features in the language vary based on the segments being concatenated? What should the default pitch contour be for this language (usually, the contour most often used with 'declarative' utterances)? What contours are used over questions? What aspects of intonation can be meaningfully varied by speakers to contribute to the overall meaning of the utterance? For example,

in tonal languages such as Mandarin, how do tones affect the overall pitch contour (e.g., is there 'tone sandhi,' or influence on the realization of one tone from a previous tone?)? Also, in languages such as Japanese, in which pitch accent is lexically specified, what sort of free prominence variation is nonetheless available to speakers? These systems must also deal with the issue of how to handle individual variation—in concatenative systems, whether to explicitly model the speaker recorded for the system or whether to derive prosodic models from other speakers' data or from abstract theoretical models. Although modeling the recorded speaker in such systems may seem the more reasonable strategy, so as to avoid the need to modify databases more than necessary in order to produce natural-sounding speech, there may not be enough speech recorded in the appropriate contexts for this speaker to support this approach, or the prosodic behavior of the speaker may not be what is desired for the TTS system. In general, though, the greater the disparity between a speaker's own prosodic behavior and the behavior modeled in the TTS system, the more difficult it is to produce natural-sounding utterances.

p0035 Whatever their prosodic inventory, different TTS systems, even those that target the same human language, will attempt to produce different types of prosodic variation, and different systems may describe the same prosodic phenomenon in different terms. This lack of uniformity often makes it difficult to compare TTS systems' capabilities. It also makes it difficult to agree on common TTS markup language conventions that can support prosodic control in speech applications, independent of the particular speech technology being employed.

p0040 TTS systems for most languages vary prosodic phrasing, although phrasing regularities of course differ by language; phrase boundaries are produced at least at the end of sentences and, for some systems, more elaborate procedures are developed for predicting sentence-internal breaks as well. Most systems developed for pitch accent languages such as English also vary prosodic prominence so that, for example, function words such as 'the' are produced with less prominence than content words such as 'cat'. The most popular models for describing and modeling these types of variation include the Edinburgh Festival Tilt system and the ToBI system, developed for different varieties of English prosody, the IPO contour stylization techniques developed for Dutch, and the Fujisaki model developed for Japanese. These models have each been adapted for other languages other their original target: Thus, there are Fujisaki models of English and ToBI models of Japanese, *inter alia*. The following section specifies prosodic phenomena in the ToBI model for illustrative purposes. The ToBi model was originally developed for standard American English; a full description of the conventions as well as training materials may be found at http://ling.ohio-state.edu/~tobi.

The ToBI System

s0015

p0045 The ToBI system consists of annotations at three time-linked levels of analysis: an 'orthographic tier' of time-aligned words; a 'break index tier' indicating degrees of junction between words, from 0 (no word boundary) to 4 (full intonational phrase boundary); and a 'tonal tier,' where pitch accents, phrase accents, and boundary tones describing targets in the fundamental frequency (f0) define prosodic phrases, following Pierrehumbert's scheme for describing AU:1 American English, with modifications. Break indices define two levels of phrasing, minor or intermediate (level 3) and major or intonational (level 4), with an associated tonal tier that describes the phrase accents and boundary tones for each level. Level 4 phrases consist of one or more level 3 phrases plus a high or low boundary tone (H% or L%) at the right edge of the phrase. Level 3 phrases consist of one or more pitch accents, aligned with the stressed syllable of lexical items, plus a phrase accent, which also may be high (H-) or low (L-). A standard declarative contour for American English, for example, ends in a low phrase accent and low boundary tone and is represented by H* L-L%; a standard yes–no question contour ends in H-H% and is represented as L* H-H%. Five types of pitch accent occur in the ToBI system defined for American English: two simple accents (H* and L*) and three complex ones (L*+H, L+H*, and H+H*). As in Pierrehumbert's system, the asterisk indicates which tone is aligned with the stressed syllable of the word bearing a complex accent. Words associated with pitch accents appear intonationally prominent to listeners and may be termed 'accented'; other words may be said to be 'deaccented.' This scheme has been used to model prosodic variation in the Bell Labs and AT&T TTS systems and also as one of several models in the Festival TTS system.

Prosodic Prominence

s0020

p0050 In many languages, human speakers tend to make content words (nouns, verbs, and modifiers) prosodic prominent or accented—typically by varying some combination of f0, intensity, and durational features—and function words (determiners and prepositions) less prominent or deaccented. Many early TTS systems relied on this simple content/function distinction as their sole prominence assignment strategy. Although this strategy may work fairly well for

short, simple sentences produced in isolation, it works less well for longer sentences and for larger stretches of text.

p0055 In many languages, particularly for longer discourses, human speakers vary prosodic prominence to indicate variation in the information status of particular items in the discourse. In English, for example, human speakers tend to accent content words when they represent items that are 'new' to the discourse, but they tend to deaccent content words that are 'old,' or given, including lexical items with the same stem as previously mentioned words. However, not all given content words are deaccented, making the relationship between the given/new distinction and the accenting decision a complex one. Given items can be accented because they are used in a contrastive sense, for reasons of focus, because they have not been mentioned recently, or other considerations.

p0060 For example, in the following text, some content words are accented but some are not:

> The **SENATE BREAKS** for **LUNCH** at **NOON**, so i **HEADED** to the **CAFETERIA** to **GET** my **STORY**. There are **SENATORS**, and there are **THIN** senators. For **SENATORS**, **LUNCH** at the cafeteria is **FREE**. For **REPORTERS**, it's not. But **CAFETERIA** food is **CAFETERIA** food.

p0065 TTS systems that attempt to model human accent decisions with respect to information status typically assume that content words that have been mentioned in the current paragraph (or some other limited stretch of text) and, possibly, words sharing a stem with such previously mentioned words should be deaccented, and that otherwise these words should be accented. However, corpus-based studies have shown that this strategy tends to deaccent many more words than human speakers would deaccent. Attempts have been made to incorporate additional information by inferring 'contrastive' environments and other factors influencing accent decisions in human speakers, such as complex nominal (sequences of nouns that may be analyzed as 'noun–noun' or as 'modifier–noun') stress patterns. Nominals such as *city **HALL*** and ***PARKING** lot* may be stressed on the left or the right side of the nominal. Although a given nominal is typically stressed in a particular way, it is a largely unsolved problem, despite some identified semantic regularities, such as the observation that room descriptions (e.g., ***DINING** room*) typically have left stress and street names (e.g., ***MAIN** Street*), although not avenues or roads, do as well. More complicating in English complex nominals is the fact that combinations of complex nominals may undergo stress shift, such that adjoining prominent items may cause one of the stresses to be shifted to an earlier syllable (e.g., ***CITY** hall* and ***PARKING** lot*).

AU:2

p0070 Other prominence decisions are less predictable from simple text analysis since they involve cases in which sentences can, in speech, be disambiguated by varying prosodic prominence in English and other languages. Such phenomena include ambiguous verb–particle/preposition constructions (e.g., *George moved behind the screen*, in which accenting *behind* triggers the verb–particle interpretation), focus-sensitive operators (e.g., *John only introduced Mary to Sue*, in which the prominence of *Mary* vs. *Sue* can favor different interpretations of the sentence), differences in pronominal reference resolution (e.g., *John call Bill a Republican and then he insulted him*, in which prominence on the pronouns can favor different resolutions of them), and differentiating between discourse markers (words such as *well* or *now* that may explicitly signal the topic structure of a discourse) and their adverbial homographs (e.g., *Now Bill is a vegetarian*). These and other cases of ambiguity disambiguable prosodicly can only be modeled in TTS by allowing users explicit control over prosody. Disambiguating such sentences by text analysis is currently beyond the range of natural language processing systems.

AU:3

Prosodic Phrasing

s0025

p0075 Prosodic phrasing decisions are important in most TTS systems. Human speakers typically 'chunk' their utterances into manageable units, producing phrase boundaries with some combination of pause, f0 change, a lessening of intensity, and often lengthening of the word preceding the phrase boundary. TTS systems that attempt to emulate natural human behavior try to produce phrase boundaries modeling such behavior in appropriate places in the input text, relying on some form of text analysis.

p0080 Intuitively, prosodic phrases divide an utterance into meaningful units of information. Variation in phrasing can change the meaning hearers assign to a sentence. For example, the interpretation of a sentence such as *Bill doesn't drink because he's unhappy* is likely to change, depending on whether it is uttered as one phrase or two. Uttered as a single phrase, with a prosodic boundary after *drink*, this sentence is commonly interpreted as conveying that Bill does indeed drink, but the cause of his drinking is not his unhappiness. Uttered as two phrases (*Bill doesn't drink—because he's unhappy*), it is more likely to convey that Bill does not drink—and that unhappiness is the reason for his abstinence. In effect, variation in phrasing in such cases in English, Spanish, and Italian, and possibly other languages, influences the scope of negation in the sentence. Prepositional phrase (PP) at-

tachment has also been correlated with prosodic phrasing: *I saw the man on the hill—with a telescope* tends to favor the verb phrase attachment, whereas *I saw the man on the hill with a telescope* tends to favor an noun phrase attachment.

p0085 Although phrase boundaries often seem to occur frequently in syntactically predictable locations such as the edges of PPs, between conjuncts, or after preposed adverbials, *inter alia*, there is no necessary relationship between prosodic phrasing and syntactic structure—although this is often claimed by more theoretical research on prosodic phrasing. Analysis of prosodic phrasing in large prosodically labeled corpora, particularly in corpora of nonlaboratory speech, shows that speakers may produce boundaries in any syntactic environment. Although some would term such behavior 'hesitation,' the assumption that phrase boundaries that occur where one does not believe they should must result from some performance difficulty is somewhat circular. In general, the data seem to support the conclusion that syntactic constituent information is one useful predictor of prosodic phrasing but that there is no one-to-one mapping between syntactic and prosodic phrasing.

s0030 Overall Contour Variation

p0090 TTS systems typically vary contour only when identifying a question in the language modeled, if that language does indeed have a characteristic 'question' contour. English TTS systems, for example, generally produce all input sentences with a falling 'declarative' contour, with only yes–no questions and occasionally sentence-internal phrases produced with some degree of rising contour. This limitation is a considerable one since most languages exhibit a much richer variety of overall contour variation. English, for example, employs contours such as the 'rise–fall–rise' contour to convey uncertainty or incredulity, the 'surprise-redundancy' contour to convey that something observable is nonetheless unexpected, the 'high-rise' question contour to elicit from a hearer whether some information is familiar to that hearer, and the 'plateau' contour ('You've already heard this') or 'continuation rise' ('There's more to come'; L-H% in ToBI) as variants of list intonation and 'downstepped' contours to convey, *inter alia*, the beginning and ending of topics.

p0095 Systems that attempt to overcome the monotony of producing the same contour over most of the text they synthesize do so in the main by allowing users to specify contour variation (as well as phrasing, accent, and other prosodic features) by using some form of markup language within the input text. Recent interest in the production of different TTS voices and of different listener impressions of the state or emotion of the speech being produced has led to some renewed interest in the relationship between pitch contour and speaker personality and speaker state. This research is still in the early stages, but thus far it has not been demonstrated empirically that speaker state (except perhaps for 'boredom'), at least, can be signaled effectively by mere contour variation: Voice quality, speaking rate, intensity, and pitch range also must be involved.

s0035 Varying Timing and Pitch Range

p0100 Two additional prosodic features that TTS systems may vary in attempting to mimic human behavior are timing and pitch range. Timing parameters may include speaking rate and pausal duration between prosodic phrases. Speakers may vary their rate to convey a different interpretation of a particular pitch contour, convey some emotional state, indicate that a phrase should be interpreted as a parenthetical remark, or convey differences in topic structure. Pitch range can also produce different 'meanings' for a given pitch contour or convey differences in discourse/topic structure. It can also indicate differences in a speaker's degree of involvement. In TTS systems, pitch range variation generally involves controlling the expansion and contraction of the overall f0 range for an utterance, parameters that may be calculated in different ways depending on the prosodic model being implemented in the system. For example, if paragraphing is used as a proxy for topic structure, systems will expand their pitch range over the first phrase of the paragraph and conclude with a phrase in a more compressed range.

s0040 Predicting Prosodic Features from Text

p0105 Once a prosodic model has been chosen for the system, and the prosodic variation that will be supported by the system, how such prosodic variation will be predicted from input text must next be determined. Early TTS systems tended to develop hand-built rule systems to predict prosody assignment based on simple part-of-speech features or more elaborate syntactic parsing. The main limitation of hand-built rule systems has been the difficulty of extending and maintaining them: New rules introduced to address perceived 'errors' in prosody assignment often have unforeseen and undesirable consequences. Also, the use of syntactic parsers has generally been limited by the limits of parser speed and accuracy.

p0110 Hand-constructed rule systems for TTS prosody assignment now have largely been superceded by corpus-based techniques, in which relatively large spoken corpora are hand labeled for prosodic features and used as training material for machine learning

algorithms, which produce decision procedures automatically based on textual analysis that can also be performed automatically. Typical features used in such corpus-based approaches include the part-of-speech of words in the input, the distance of words from the beginning and end of an utterance, and so on. Automatically derived decision procedures are typically limited by the amount of hand-labeled data available for training, which has led to various attempts at speeding up hand labeling, or dispensing with it altogether, by asking native speakers to assign the prosody they might use if they uttered the sentence they are being asked to label to that sentence. Automatically derived procedures are also difficult to correct, if 'errors' in their assignment are identified, without simply providing 'correct' examples in sufficient quantity in the training corpus to outweigh the data that have led to undesirable prediction.

p0115 Machine learning approaches to accent prediction for TTS have typically used automatic part-of-speech tags (often including the part-of-speech of items in a window around the word whose accent status is to be predicted), distance of the word from the beginning or end of the sentence, and location of punctuation near the word to achieve fairly good performance ($\geq$70% accuracy when tested on labeled human data). More ambitious systems have attempted to identify the word's information status (given or new, contrastive or not) or special constructions, such as complex nominals, preposed modifiers, and discourse markers. Some recent success has been had by including word frequency and mutual information scores into the mix. Still more ambitious systems have been proposed for concept (message)-to-speech systems, in which syntactic, semantic, and discourse information is available during the generation process. However, to date these have not improved over prosodic assignment in text-based systems, largely because the correlation between prominence and other prosodic features and other linguistic information is still not well enough understood to permit more accurate prediction.

p0120 Like prominence assignment procedures, phrasing assignment in TTS systems is done either by rule or by corpus-trained procedures. Early phrase prediction in TTS systems relied on spotting key words or punctuation to determine prosodic boundary location: So, this sentence would be produced by a TTS system as two prosodic phrases due to the presence of a comma after 'so'. However, text punctuation is too sparse and often too erratic to rely on entirely for the insertion of natural-sounding phrase boundaries, especially for very long sentences. Also, some punctuation should not signal a major phrase boundary (e.g., *He comes from Little Rock, Ark.*). More sophisticated phrasing prediction procedures have been developed by hand or via machine learning techniques, as discussed previously for prominence decisions. Hand-built rule systems tend to rely on syntactic information produced by a parser, whereas automatically trained techniques tend to rely on simple part-of-speech windows, sentence length, distance of the potential phrase boundary from the beginning and end of the sentence, sentence type (question or declarative), and punctuation. However, some success has been had in recent years by incorporating more sophisticated constituent information together with these simpler features. The intuition between the correlation of prosodic phrasing and syntactic structure is that boundaries may tend not to occur internal to certain constituents and to divide other constituents from one another. Dynamic information about prior prosodic decisions also represents an important area for future study since, for example, it is less likely that a phrase boundary will appear very soon after a previous boundary; similarly, corpus-based analysis has shown that phrase boundaries occur rarely between two deaccented words. State-of-the-art performance in prosodic phrase prediction is generally reported to be in the low to mid–90s for precision compared to a labeled test corpus.

User-Specified Control of Prosodic Variation

s0045

p0125 With some degree of explicit user control over prosodic variation, TTS can sound much more natural. This control is accomplished by providing explicit markup capabilities such that users may specify, by script or by hand, how various prosodic parameters should be set for any input text. Systems currently provide this capability via system-specific markup languages. However, markup standards are being adopted gradually that permit some general forms of prosodic control. The most popular of these currently appears to be SABLE, which has been implemented in the Festival TTS system and in the Bell Labs TTS system. Earlier markup languages that contributed to the development of SABLE were JSML (Java Speech Markup Language), SSML (Speech Synthesis Markup Language), and STML (Spoken Text Markup Language). The main limitation of these languages is that they must be mapped to the proprietary languages that existing systems provide, which commonly provide greater prosodic control that the conventional markup language. Until systems begin to be developed for standard markup languages (and until these markup languages include more fine-grained control over prosodic variation, which will involve major agreements on how to de-

fine and specify that variation), this dichotomy will remain a problem.

p0130 There is considerable practical wisdom indicating that if a TTS system is targeted at a particular application, prosodic and other features can be tuned to this application with good effect. For example, a NYNEX experiment involving a reverse telephone directory showed that the worst performing TTS system could become the most successful by modeling the prosodic behavior of the human attendants whose role was being assumed by the TTS system. In general, if the input to a TTS system is characterizable in advance (e.g., names and addresses, automobile reservations, and a financial domain) scripts can be devised to use these assumptions to improve performance.

Evaluation

s0050

p0135 Evaluating TTS systems in general is extremely difficult. Now that most systems are quite intelligible, most testing centers on their naturalness, which is both difficult to define for raters and difficult to attribute to different components of the system. Evaluating prosodic assignment has generally been done either through subjective judgment ratings, such that the same sentence is presented in its original form with some less sophisticated engine producing prosodic assignments, or through some new and improved assignment algorithm. Unfortunately, prosodic assignment may interact in unpredictable ways with other components of the system: For example, if the prosodic assignment component correctly specifies that a yes–no question contour be used but the database cannot produce such a contour, the prosodic assignment component may be poorly evaluated in such tests. Assigning 'blame' to various components is in general quite difficult since small failures in components may contribute to a lack of naturalness.

AU:4

p0140 A more objective, but still flawed, approach is to evaluate the prosodic assignment procedure's automatic specification on a hand-labeled test set of human speech. There are always multiple ways of producing any utterance, however. Thus, comparison of TTS output to any single prediction would appear to be too strict a metric. However, it is not clear what kind of comparison to human performance would be more valid. Some have compared their prosodic assignment to the combined production of multiple speakers, but this has clear drawbacks. For example, allowing a phrase boundary in a given sentence wherever one of multiple human speakers produces a boundary can easily result in a sentence with far too many boundaries for any human production. Also, the location of one boundary may depend on the presence of another, so a production including only one may sound distinctly odd. Task-based evaluations have also been attempted with some success, as in the reverse telephone directory evaluation mentioned previously, in which subjects were asked to transcribe the names and addresses the system produced; a TTS system with worse overall intelligibility performed better than other systems when its prosody was changed to model that of the human attendant. However, this modification largely involved simply slowing the rate of the synthesizer. In summary, the evaluation of TTS systems in general, and of prosodic assignment in particular, remains a major research question.

Conclusion

s0055

p0145 In the current age of corpus-based TTS systems that rely on searching a large inventory of speech for variable-length units to concatenate, with relatively little system modification of prosodic features, effective and realizable prosodic assignment represents a major problem. There are many issues to be resolved regarding how natural-sounding prosody can be assigned, based on text input or even on more information-rich representations. However, if such prosodic variation is not actually possible within a TTS system without degrading the overall quality of the modified speech, the field has a serious problem. Therefore, not only must there be research on how to assign prosodic features more effectively (particularly full pitch contours) but also the overall architecture of the system must make it possible to realize such prosodic assignment in the output speech. Some very natural-sounding systems still cannot produce yes–no question intonation reliably, for example, because too few question contours occur in the system database.

See also:

AU:5

Bibliography

Damper R I (ed.) (2001). *Data-driven techniques in speech synthesis*.

AU:6

Jun S A (ed.) (2004). *Prosodic models and transcription: towards prosodic typology*. Oxford: Oxford University Press.

Sagisaka Y, Campbell N & Higuchi N (eds.) (1996). *Computing prosody: computational models for processing spontaneous speech*. New York: Springer-Verlag.

Sproat R (1998). *Multilingual text-to-speech synthesis: the Bell Labs approach*. Boston: Kluwer.

Van Santen J P H, Sproat R W, Olive J P *et al.* (eds.) (1996). *Progress in speech synthesis*. New York: Springer-Verlag.

Article Number: LALI: 00914

Author Query Form

Book: Encyclopedia of Language and Linguistics
Article No.: 00914

Dear Author,

During the preparation of your manuscript for typesetting some questions have arisen. These are listed below. Please check your typeset proof carefully and mark any corrections in the margin of the proof or compile them as a separate list. Your responses to these questions should be returned within seven days, by email, to Dr Jennifer Lai, email: lai@watson.ibm.com, and copied to MRW Production, email: lali_proofs@elsevier.com

Query	Details Required	Author's response
AU1	''annotations at three time-linked levels of analysis: an 'orthographic tier' of time-aligned words; a 'break index tier' indicating degrees of junction between words, from 0 (no word boundary) to 4 (full intonational phrase boundary); and a 'tonal tier,' where pitch accents, phrase accents, and boundary tones describing targets in the fundamental frequency (f0) define prosodic phrases, following Pierrehumbert's scheme for describing American English, with modifications'' as meant? There are only 3 levels of analysis listed, rather than 4 (as in original sentence).	
AU2	''Although a given nominal is typically stressed in a particular way, it is a largely unsolved problem'' as meant?	
AU3	''cases of ambiguity disambiguable prosodicly'' as meant?	
AU4	''Evaluating prosodic assignment has generally been done either through subjective judgment ratings, such that the same sentence is presented in its original form with some less sophisticated engine producing prosodic assignments, or through some new and improved assignment algorithm'' as meant?	
AU5	Pls provide Cross-References.	
AU6	Pls provide publisher and location for Damper (2001).	
AU7	Pls provide brief author biography and keywords.	

Article Number: LALI: 00914

Non-Print Items

Abstract:

Text-to-speech (TTS) systems face two problems in modeling prosody: (1) determining an appropriate prosody for input text ('How would a person say this sentence?') and (2) realizing the prosody so specified in a natural-sounding way. Failures in these two areas are a major reason why TTS systems still often sound unnatural, particularly when they are producing larger stretches of output. This article describes the prosodic phenomena TTS systems typically try to model and some of the barriers to modeling these successfully.

AU:7 ### Biography:

AU:7 **Keywords:**

Author Contact Information:

Julia Hirschberg
Department of Computer Science
University of Columbia
1214 Amsterdam Avenue, M/C 0401, 450 CS Building
New York, NY 10027
USA
julia@cs.columbia.edu