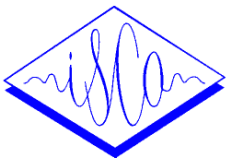

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

A CORPUS-BASED SYNTHESIZER

Richard Sproat
Julia Hirschberg
David Yarowsky

Linguistics Research Department
AT&T Bell Laboratories
Murray Hill, NJ, 07974, U.S.A.

ABSTRACT

This paper describes NewExpress, the new text-to-phonetic-representation component of the AT&T Bell Laboratories Text-to-Speech system (TTS). To the best of our knowledge, NewExpress represents the first extensive use of corpus-based linguistic techniques in a text-to-speech program. We discuss the use of such techniques in the system in four main areas: general pitch accent assignment, prosodic phrasing, pitch accent assignment in noun compounds, and homograph disambiguation. We demonstrate that these techniques afford an improvement in the performance of TTS.

10.21437/ICSLP.1992-130

1 INTRODUCTION

This paper describes three applications of corpus-based linguistic techniques implemented in NewExpress, the new text-to-phonetic-representation component of the AT&T Bell Laboratories Text-to-Speech system. As far as we know, NewExpress represents the first extensive use of such methods within a text-to-speech system. Below, we describe applications of corpus-based methods to four major problems in text-to-speech: general pitch accent assignment, the determination of prosodic phrasing, accent assignment within noun compounds, and homograph disambiguation.

2 PITCH ACCENT AND PHRASING PREDICTION

The association between prosodic variations and semantic, syntactic and discourse features of utterances has long been an important issue in theoretical studies of language as well as in applications to speech synthesis. Most current text-to-speech systems use simple word-class information to assign pitch accent: *function words*, such as prepositions, are deaccented while *content words*, such as nouns, are accented. Intonational boundaries are placed where non-final punctuation occurs in text or according to very simple parsing of the input text. Message-to-speech systems and text-to-speech systems for restricted domains take advantage of richer semantic, syntactic, and discourse-level information [18, 5, 11], but such information is not available for unrestricted text-to-speech. While truly natural prosodic assignment is not currently achievable for text-to-speech, current corpus-based analysis techniques applied to relatively large prosodically labeled corpora, do make it is possible to improve prosodic assignment considerably with fairly simple information inferred from unrestricted input.

Such information is being exploited for the assignment of prosodic features in NewExpress. Both phrasing and accent algorithms are derived from the analysis of sizeable, prosodically labeled corpora. For pitch accent placement, these included single and multi-speaker corpora of radio speech, multi-speaker spontaneous (elicited) speech, and single speaker read sentences; phrasing prediction procedures have been developed both from spontaneous and from read multi-speaker speech.

2.1 Pitch accent prediction

Models for accent placement were derived both by hand and automatically from corpus analysis. The hand-crafted rules model observed accent decisions on radio speech test sets with 82.4% accuracy; on citation-format utterances, they performed at 98.3% accuracy. Procedures developed automatically produce decision trees which predict radio speech accent with 85.1% success. These trees were produced via Classification and Regression Tree analysis (CART) [1] techniques.[1]

The rules currently implemented in TTS make use of part-of-speech and morphological information to assign input tokens to one of four broad classes — closed-cliticized, closed-deaccented, closed-accented, and open — based upon frequency distributions in the training data. For each token the following additional information is collected: preposed adverbials are identifed from surface position and part-of-speech, as are fronted *PP*s, and labeled as potentially *contrastive*. *Cue phrases* (discourse markers, such as 'well' and 'now' which provide explicit structural information about the text) are identified from surface position and part-of-speech, and their accent status is predicted following findings in [13]. Verb-particle constructions are identified by table look-up. Local focus is implemented as a stack of lemmas of all content words in a phrase. New items are pushed on the stack as each phrase is read and subsequently treated as 'given', and thus potentially deaccentable. Cue phrases trigger either push or pop operations, roughly as described in [9]. Paragraph boundaries cause the entire stack to be popped. Noun compounds and their citation-form stress assignment are identified by the NP component described in Section 3. Finally, possible contrastiveness is inferred by comparing the presence of roots of elements of a nominal in local focus; if some items are 'given' and others 'new', the new items are marked as potentially contrastive. Accent assignment is then determined as follows: assign 'closed-cliticized' and 'closed-deaccented' items the status accorded their class. Next, 'contrastive' items are assigned emphatic accent. Then 'closed-accented' items are accented, remaining 'given' items deaccented, remaining noun-compound elements assigned their citation-form

[1]Success rates cited here are cross-validated (determined by successive training on 90% of the data and testing on 10%, averaging the results) from feature values available automatically for text-to-speech.

accent, and all other items accented.

2.2 Phrase boundary prediction

Phrasing procedures are derived using CART techniques on a sample from the DARPA ATIS corpus [6]. Resulting prediction trees achieve cross-validated success rates on multi-speaker read speech of 91.2% for major boundaries only, 88.4% for collapsed major/minor phrases, and 81.9% for a three-way distinction between major, minor and null boundary. For spontaneous speech, they predict correctly in 88.2% of cases for major only, 84.4% for collapsed major/minor phrases, and 78.9% for the three-way distinction.

The current prototype boundary assignment module identifies phrase boundaries based upon decision trees trained on read speech only. The following automatically-available variables were examined for each potential boundary location $< w_i, w_j >$ (where w_i represents the word to the left of the site and w_j the word to the right): temporal information (including length of utterance in words and seconds, rate, distance of w_i in syllables, lexically stressed syllables, words, and seconds from the beginning of the utterance and distance of w_i in words and seconds from the end utterance); predicted pitch accent (as described in Section 2.1) for w_i and w_j and stress level of last syllable in w_i; part-of-speech for a four word window around $< w_i, w_j >$; largest syntactic constituent dominating w_i but not w_j and vice versa, and smallest constituent dominating them both; whether $< w_i, w_j >$ is dominated by an NP and, if so, distance of w_i from the beginning of that NP, the NP, and distance/length; and mutual information scores for a four-word window around $< w_i, w_j >$. The most successful of these predictors so far appear to be part-of-speech, some constituency information, and mutual information; each can predict a large percentage of observed boundaries. However, more training data is clearly needed to improve predictive power, particular upon text not well represented by the training corpus.

Since the current method of obtaining training data (prosodic labeling of a large corpus) is fairly slow and labor intensive, an alternate method for augmenting the original corpus is being tested. First, text is chosen at random from the AP news wire and phrasing predictions are obtained from the corpus-based prediction trees. Next, these results are corrected by hand to insert or delete undesired phrase boundaries, where desirability is determined by human subjects. Finally, these sentences are treated as data for the training of new phrasing predictors.

3 NOUN-NOUN COMPOUND ACCENT

Sequences of nouns in English text — *noun compounds* — pose a difficult problem for text-to-speech systems, because lexical accent can in principle be assigned to any member [14, 16].[2] Thus — failing pragmatic reasons to do otherwise — English speakers would generally accent the first word (and deaccent the second) in *ARBITRATION panel*, but accent the second in *city LIFEGUARD*. It is often argued that accent placement depends partly upon the semantic relationship between the words in the compound, and in part upon purely lexical factors [7, 12]. Indeed, the compound analyzer in NewExpress — NP — uses a database of literal compounds listed with their accent attributes (e.g. *littleneck CLAM*), plus a database of quasi-semantic rules to predict accent: so a rule of the form *ROOM + HOUSEWARE* ⇒ *righthand accent* derives the correct accent for *kitchen PHONE*. This system is described in detail in [16]. While this approach fares well enough on material that it can sensibly analyze, one still encounters a large number of noun-noun compounds for which the system gives no analysis, and in those cases it reverts to the default of assigning accent to the penultimate member of the sequence. Since this fallback option is only right about 75% of the time in most speakers' judgment, it is not really satisfactory.

As a supplement to the rule-based methods used in NP, we have trained a simple statistical method on a hand-labeled corpus of 7831 noun compound types picked randomly from the 1990 Associated Press newswire, and tagged for lefthand or righthand accent by one of the authors. Of these, 88% (6891) compounds were used as a training set and the rest were set aside as the test set. For each compound in the training set and for each of the two lemmata in the compound we collect the set of broad topical categories associated with the lemma in *Roget's Thesaurus* [3]. Then, for each element of the cross-product of these Roget categories we tally the accent (left or right) tagged for the compound. Intuitively, the cross-product of the categories gives a crude encoding of the set of possible semantic relationships between the two words. We also create entries for the first lemma (modifier) *qua* modifier, and the second lemma (head) *qua* head, and tally the accent for those entries. For a compound in the test set, we sum the accent pattern evidence accumulated for each of the elements in the cross product of the Roget categories for the compound, and the head/modifier entries for the two lemmata, selecting the accenting that wins. As an example, consider *cloth DIAPERS*, which occurs in the test set and not in the training set. In the training set, the lemma *diaper* never occurs in the righthand position; the lemma *cloth* occurs twice in the lefthand position, once in a compound with lefthand accent (*CLOTH merchant*) and once in a compound with righthand accent (*cloth BANNERS*). This amounts to no evidence, so one would by default assign lefthand accent to *cloth diapers*; however this compound also matches the category sequence MATERIALS/CLOTHING, for which there is a score of 13 favoring righthand accent. In practice it was found that a large number of the Roget cross-categorial combinations were not useful and often in fact detrimental: this is because Roget categories are usually topical rather than taxonomic classifications. However, 19 fairly taxonomic combinations were retained.

The overall results of the experiment were as follows. For the 940 nominals in the test set there was an error rate of 16%: note that the error rate for uniformly assigning lefthand accent is 30% for this set, so the approach roughly halves the error rate of the fallback option. (Of the 30% in the test set that were hand-tagged with righthand accent, 58% were correctly assigned righthand accent.) Only 35 nominals in the test set matched one of the pruned set of cross-categorial combinations. However, of these 27 (77%) correctly predicted accent on the basis of the category combinations alone. This suggests that a more reasonable taxonomic categorization than *Roget's* could be useful in accent assignment to nominals.

At present, the statistical method is used within NP as a backup for noun-noun sequences in case the input cannot be analyzed by NP's rule-based methods. To evaluate the system, we had six judges independently judge accent placement for 1138 compound *types* picked randomly from the *1992* Associated Press newswire. Pairwise similarity measures between judges averages 0.91 and this may be taken as an upper bound for performance of computer models. The baseline algorithm of always assigning lefthand ac-

[2] Although the NP component of NewExpress analyzes compounds of length greater than two, we will restrict our discussion here to binary cases, which comprise the majority of examples (90%) of noun compounds that one encounters in text.

cent agrees with the judges on average at 0.76, whereas NP agrees at 0.85, meaning that NP covers about 67% of the difference in performance between humans and the baseline. If one counts by token rather than by type, then the judges agree at 0.93, and the baseline/judge and NP/judge agreement scores are respectively 0.74 and 0.89. Interestingly, only 174 (15%) of the cases were actually handled by the hand-built rules of NP: the 964 remaining cases were assigned accent on the basis of the trained models described above, suggesting that corpus-trainable models are more extensible than hand-built systems in this domain.

4 HOMOGRAPH DISAMBIGUATION

The processing of words with multiple pronunciations is an instance of the word sense disambiguation problem, and we have employed statistical techniques developed for this larger task in the new pronunciation rules for TTS. Homographs are of three types: I) *Part of Speech:* Ambiguities such as *lives* [lɪvz / laɪvz] or *read* [ri:d / rɛd] are handled by a stochastic tagger [4] augmented with word-specific optimizations which substantially reduce its error rate. II) *Capitalization:* Proper names such as *Nice* and *Begin* are ambiguous in certain contexts, such as sentence initial position, titles and single-case text. III) *Polysemy within part of speech:* Words including *bass* and *bow* require additional "semantic" evidence for disambiguation.

Our previous work in sense disambiguation [8, 17] has been based on wide-context Bayesian discrimination, where words in an n-word window independently contribute positive or negative evidence for a given sense. Due to this strict independence assumption, the method cannot exploit evidence conditional on other evidence. For example, *take* is evidence for the non-metallic sense of *lead* only in collocations such as *take the lead*, a condition which cannot be expressed in this formalism. Others [2] have used classification trees, but due to the conditional branching at each tree level, have encountered problems with sparse-data estimation in a very large parameter space.

Our new approach combines the strengths of these two methods, by using an n-gram model to capture local, conditional dependencies, and uses probabilities derived from a wide-context Bayesian model to capture long-distance semantic collocations. As in [10], the system incorporates several potential sources of evidence, such as words, parts-of-speech, and lemmas in specific positions, as well as questions about the ambiguous word (e.g., is it capitalized?). The strength of each piece of evidence (E_i) is expressed in log likelihoods [15], $log_2\ Pr(E_i|Pronunciation_A)\ /\ Pr(E_i|Pronunciation_B)$. Traditionally the log likelihoods are summed. However, there are problems due to the non-independence in the multiple sources of evidence, and because one count in the ratio is often zero, the result is highly sensitive to the smoothing strategy used. Rather than combining all the available evidence probabilities, we discovered that performance actually improves when only the single strongest piece of evidence is used.[3] Thus we use the best evidence first, sorting by log likelihood and using the first pattern that matches. Sample, abbreviated decision tables are outlined below:

Decision Table for *wound*

Logprob	Position	Type	Evidence	Pronunciation
14.13	+1	WORD	*up*	⇒ waʊnd
12.01	−1	PT.OF.SP	<ARTICLE>	⇒ wund
11.95	−1	WORD	*gunshot*	⇒ wund
11.23	+1	PT.OF.SP	<SUBJ-PRO>	⇒ waʊnd
10.70	−1	WORD	*bullet*	⇒ wund
9.72	−1	LEMMA	*have/V*	⇒ waʊnd
9.70	−1	WORD	*head*	⇒ wund
9.45	IN_SENTENCE	LEMMA	*coil*	⇒ waʊnd

Decision Table for *putting*

Logprob	Position	Type	Evidence	Pronunciation
12.82	+1	LEMMA	*green/N*	⇒ pʌtɪŋ
11.32	+1	LEMMA	*surface/N*	⇒ pʌtɪŋ
10.40	IN_SENTENCE	LEMMA	*wedge*	⇒ pʌtɪŋ
10.25	IN_SENTENCE	LEMMA	*golfer*	⇒ pʌtɪŋ
8.89	IN_SENTENCE	LEMMA	*golf*	⇒ pʌtɪŋ
		DEFAULT		⇒ pʊtɪŋ

Rule sets constructed in this way will initially exhibit considerable redundancy. The first type, redundancy by subsumption, is a consequence of a weighting scheme favoring the most general, unambiguous statement possible. Thus a lemma with high log likelihood will eclipse its member words, and an unambiguous bigram will eclipse any dependent trigrams. These are easy to identify and remove; higher order n-grams are not even generated if the lower-order form is unambiguous. A more subtle case is redundancy by association. For example, *riot, soldier*, and *demonstration* are all strong indicators of the crying sense of *tear* in wide context, but only because of their exclusive association with the stronger adjacent indicator tear *gas*. When treated as independent probabilities in a traditional Bayesian framework, these highly correlated sets of words often yield grossly over-inflated confidence scores. These particular words contribute little to the classification of novel contexts as they almost never appear independent of *gas*, and when they do there is no evidence that they are more indicative of the crying sense of tear than the ripping sense.

We employ several techniques for removing both sources of redundancy. The simplest is to train a discriminator and then apply it directly to the training set. Count the number of times each rule is the first match for a training example, decrementing in cases where the use yields an incorrect classification. Retain the rules which actually contribute to the modeling of the training data, in their original order. Some words pruned in this way may have contributed to the classification of testing examples. A 3% drop in performance is observed when all redundancy by association is removed, but an over 90% reduction in space is realized. The optimum pruning is subject to cost-benefit analysis.

Training material is acquired through an iterative bootstrapping procedure. Uniform lack of ambiguity in collocations is a useful way of identifying probable tagging errors. The base set is often derived from our class-based sense disambiguator using Roget's Thesaurus [17]. This method offers full vocabulary coverage with no hand-tagging, but at the cost of reduced precision. For the current system, we have made an investment in partial hand-tagging to achieve improved precision relative to more fully self-organizing methods.

[3] A possible explanation for this is based on another interesting discovery – that collocations between content words (Noun, Verb, Adverb, Adjective) are overwhelmingly unambiguous. In a sample study of adjacent collocations with content words, pronunciations were ambiguous in fewer than 1% of the bigrams with frequency ≥ 3, and fewer than 0.1% of the bigrams when weighted by token. The very rare exceptions include *fish for* [*black bass*] / *a* [*black bass*] *player*. It appears that a single observed bigram with a content word is enough to motivate that pronunciation for future instances of the bigram, not only with the given word but with its lemma. The strength of this property decreases with distance, but remains very strong for collocations within a ±3 word window.

Homograph Disambiguation: System Performance					
Word	Pron1	Pron2	Sample Size	Prior Prob.	System Performance
lives	laɪvz	lɪvz	33186	.69	.98
wound	waʊnd	wund	4483	.55	.98
Nice	naɪs	nis	573	.56	.94
Begin	brgɪn	beɪgɪn	1143	.75	.97
Chi	tʃi	kaɪ	1288	.53	.98
Colon	koʊloʊn	ˈkoʊlən	1984	.69	.98
lead (N)	lid	lɛd	12165	.66	.98
tear (N)	tɛɚ	tɪɚ	2271	.88	.97
axes (N)	ˈæksiz	ˈæksɪz	1344	.72	.96
IV	aɪ vi	fɔɪθ	1442	.76	.98
Jan	dʒæn	jan	1327	.90	.98
routed	ɹutɪd	ɹaʊtɪd	589	.60	.94
bass	beɪs	bæs	1865	.57	.99
TOTAL			63660	.67	.97

Performance on type I (part of speech) ambiguities is best measured in terms of improvement over existing taggers. Typically, there is a baseline of roughly 80% achievable by a small set of almost unambiguous part-of-speech sequences (e.g. Det (N|V), (N|V) Modal). For the remaining difficult cases, performance of the existing tagger is often only slightly better than chance. By incorporating lexical collocational information as well, we can disambiguate examples such as "*a bullet* WOUND *under his*" and "*the cable* WOUND *around the*" which are not distinguished by part-of-speech sequence alone. This increases tagger performance for *wound* from 82% to 98%. Overall we observe over 60% reduction in error rate, yielding a mean precision of 97%.

Performance on a sample of type II and III ambiguities (capitalization and within part of speech) is outlined in the preceding table. All results are based on 5-fold iterative cross validation. Performance depends on the window size used. If only immediately adjacent (±1 word) context is examined, mean precision is 92%. Using only a ±3 word window yields 94% precision, and allowing examination of the full sentence results in system performance of 97%. Use of broader discourse context has been shown to improve performance further, and may be productively utilized in the future.

5 CONCLUSION

In this paper we have described applications of corpus-based techniques to the development of new procedures for word pronunciation, pitch accent assignment, and prosodic phrasing in TTS. These procedures demonstrate that the distributional properties of language can be employed to improve the naturalness of text-to-speech synthesis.

References

[1] Leo Brieman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees.* Wadsworth & Brooks, Monterrey CA, 1984.

[2] Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. Word sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting*, pages 264–270, Morristown, NJ, 1991. Association for Computational Linguistics.

[3] Robert Chapman. *Roget's International Thesaurus.* Harper and Row, New York, fourth edition, 1977.

[4] Kenneth Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Morristown, NJ, 1988. Association for Computational Linguistics.

[5] Laurence Danlos, Eric LaPorte, and Francoise Emerard. Synthesis of spoken messages from semantic representations. In *Proceedings of the 11th International Conference on Computational Linguistics*, pages 599–604. International Conference on Computational Linguistics, 1986.

[6] DARPA. *Proceedings of the DARPA Speech and Natural Language Workshop*, Hidden Valley PA, June 1990.

[7] Eric Fudge. *English Word-Stress.* Allen and Unwin, London, 1984.

[8] William Gale, Kenneth Church, and David Yarowsky. A method for disambiguating word senses in a large corpus. *Computers and Humanities*, 1992.

[9] Barbara Grosz and Candace Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[10] Marti Hearst. Noun homograph disambiguation using local context in large text corpora. In *Using Corpora*, Waterloo, Ontario, 1991. University of Waterloo.

[11] Jill House and Nick Youd. Contextually appropriate intonation in speech synthesis. In *Proceedings of the ESCA Workshop on Speech Synthesis*, pages 185–188, Autrans, 1990. ESCA.

[12] Mark Liberman and Richard Sproat. The stress and structure of modified noun phrases in English. In Anna Szabolcsi and Ivan Sag, editors, *Lexical Matters.* CSLI (University of Chicago Press), 1992.

[13] Diane Litman and Julia Hirschberg. Disambiguating cue phrases in text and speech. In *Proceedings of COLING90*, Helsinki, August 1990. COLING.

[14] Alex Monaghan. Rhythm and stress-shift in speech synthesis. *Computer Speech and Language*, 4(1):71–78, 1990.

[15] Fredrick Mosteller and David Wallace. *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, Reading, Massachusetts, 1964.

[16] Richard Sproat. Stress assignment in complex nominals for english text-to-speech. In *Proceedings of the ESCA Workshop on Speech Synthesis*, pages 129–132, 1990.

[17] David Yarowsky. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proceedings of COLING-92*, Nantes, France, July 1992. COLING.

[18] S. J. Young and Frank Fallside. Speech synthesis from concept: A method for speech output from information systems. *Journal of the Acoustic Society of America*, 66(3):685–695, September 1979.