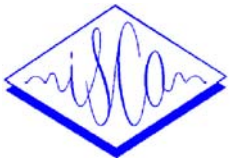

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

DESIGN AND COLLECTION OF A CORPUS OF POLYPHONES AND PROSODIC CONTEXTS FOR SPEECH SYNTHESIS RESEARCH AND DEVELOPMENT

Kim Silverman,[1] *Victoria Anderson,*[2] *Jerome Bellegarda,*[1] *Kevin Lenzo,*[3] *and Devang Naik*[1]

[1] Apple Computer, Cupertino, California 95014, USA
[2] fonix Corporation, Cupertino, California 95014, USA
[3] Carnegie-Mellon University, Pittsburg, Pennsylvania 35069, USA

ABSTRACT

The design principles and collection procedures behind a speech synthesis corpus directly impact the performance of the resulting text-to-speech system. This paper describes the design and collection of the Victoria corpus, created to support speech synthesis research and development at Apple Computer. This corpus is composed of five constituent parts, each designed to cover a specific aspect of speech synthesis: polyphones, prosodic contexts, reiterant speech, function word sequences, and continuous speech. It was spoken in general U.S. English by one linguistically-trained adult female. Portions of the corpus are being used in the statistical estimation of duration and pitch models for Apple's next-generation text-to-speech system, MacinTalk 4.

1 INTRODUCTION

In recent years, text-to-speech (TTS) systems have come to rely more and more on speech corpora. One reason has been the emergence of concatenative synthesis, which requires recorded speech for unit selection. Another factor has been the steady shift from hand-written, hand-tuned prosody rules to formal multi-variate prosodic models, which requires the associated model parameters to be statistically derived from a training corpus. When the prosodic phonological markup is used to control more than the traditional pitch and duration, such as the characteristics of the glottal excitation function and overall spectral slope, a greater number of parameters must be derived from this corpus. In the future, other aspects of prosody will likely be investigated, such as the differences between citation forms and connected speech, or the problem of paragraph-length prosody. As a result, even more parameters will need to be derived from the corpus.

In statistical modeling, the model parameters are typically estimated from a large amount of carefully collected data. The reliability of this estimation depends critically on the quantity, coverage, and consistency of the data available. As a result, the design principles and collection procedures behind the training corpus have a direct impact on the quality, communicative effectiveness, and naturalness of synthetic speech.

Beyond the minimal requirement that statistical models be estimated on "enough data," there is no standard approach to corpus design. Corpora vary from systematically-generated nonsense words, through lists of discrete unrelated sentences, to news broadcasts. Some contain a single speaker, others span multiple speakers. This variety can be traced to divergent goals in data collection. Is the purpose just to provide speech synthesis units? Or to also provide data for studies in methods of signal representation? Should the same speaker be used to construct both acoustic and prosodic models? Etc...

This paper describes the design and collection of a large corpus, informally referred to as the Victoria corpus, created to support speech synthesis research and development at Apple Computer. This corpus is composed of five major sub-corpora, each designed to cover a specific aspect of speech synthesis: polyphones, prosodic contexts, reiterant speech, function word sequences, and continuous speech. It was spoken in general U.S. English by one linguistically-trained adult female (VA). Portions of the corpus are being used in the statistical estimation of duration and pitch models for Apple's next-generation text-to-speech system, MacinTalk 4.

The paper is organized as follows. In the next section we give an overview of the corpus and its various constituent parts. Section 3 identifies some of the important procedural issues to be addressed in collecting this kind of corpus. Finally, Section 4 describes in greater detail the portion of the corpus which we have used so far for statistical duration modeling.

2 CORPUS DESIGN

The Victoria corpus is intended to: (i) support research into high-quality signal representation, (ii)

provide source units for concatenative synthesis, and (iii) be of sufficient quality to support detailed pitch extraction, pitch epoch detection, and inverse filtering for estimation of glottal source parameters. Following is a brief overview of each of its five constituent parts.

2.1 Polyphones

This segment of the corpus was designed to provide various types of units for concatenative synthesis, such as common syllables or polysyllabic strings, words, common word stems, and/or common inflectional morphemes. It uses the most common 28,000 words of U.S. English, as estimated from several text corpora available from the LDC [1]. These words were arranged into pairs, which were in turn placed in sets of 2 or 3 (separated by commas) to form "sentences," using a rich sampling of phoneme concatenations as the phonetic criteria for word groups. Half of these sentences ended with a period, and half with a question mark. A representative example is:

rife undertaking, anyhow fanatic, grab disruptive? (1)

Each word was thus spoken with four distinct intonational patterns (cf. [2]): (i) **H* L L**, a sentence-final falling nuclear accent, as if followed by a period; (ii) **H* L H**, a nuclear falling-rising pattern that typically occurs before a comma; (iii) **H* H H**, a nuclear high-rise that is common in unmarked yes-no questions and (iv) **H***, a prenuclear high pitch accent. The end result is that each of the 28,000 words was produced in four utterance positions: (i) utterance-initial, (ii) phrase-initial but utterance-medial, (iii) phrase-final but utterance-medial, and (iv) utterance-final. These by no means exhaustively represent the rich variation used in normal everyday conversation, but they cover the most common and perceptually-salient subset in the informative factual discourse to which synthetic speech is often applied.

2.2 Prosodic Contexts

This segment of the corpus was designed to systematically cover all syllable shapes with rich phonetic variation. It complements the above segment, in that it comprises the most common prosodic boundaries in English which are not represented at all in the Polyphones sub-corpus. Here we systematically vary the distance between pitch accents, and between accents and the next prosodic boundary. More details are given in Section 4.

To also support studies in speaking rate variation, we have recorded one pass through this sub-corpus at the speaker's fastest possible speaking rate. Note that we did not attempt to repeat this exercise at the speaker's slowest possible (!) speaking rate. This was for two reasons: (i) there is little use for slow speech in speech synthesis applications, and (ii) we believe that most instances of slow speech, where speakers and listeners perceive that the speaking rate has slowed down to increase clarity, actually consists of more frequent and more major prosodic boundaries.

2.3 Reiterant Speech

This segment of the corpus was designed to support detailed modeling of pitch contours for the intonations in the above two segments, but uncontaminated by segmental perturbations. Semantically-coherent utterances were constructed to cover a superset of the intonational melodies in the above sub-corpora, on a subset of the syllable structures. Each utterance was spoken with the intended intonation, then immediately followed by a version where every open syllable was replaced with "ma" and every closed syllable replaced with "mom."

This yields very smooth and reliable pitch contours, which enables accurate measures of the fit of parameterized pitch models using gradient descent and multiple regression. The resultant pitch model for MacinTalk 4, and some implications for prosodic phonology, will be described in a subsequent paper.

2.4 Function Word Sequences

This segment of the corpus was designed to cover frequent sequences of unstressed function words (such as "and he has," "in that," and "that we have") and some clitic groups (such as "couldn't've"). These are notoriously difficult to synthesize, extremely common in connected speech, yet totally lacking in the above sub-corpora. The list of function word sequences was automatically derived from the Wall Street Journal corpus [3].

2.5 Continuous Speech

In all prior segments, utterances are in citation form, which runs the danger of producing over-articulated synthetic speech. Besides, citation forms are inherently inadequate to model larger-span effects, such as prosodic behavior in very long sentences, paragraph-length prosody [4], and the preferred speaking rate for communicative connected speech.

This segment of the corpus was designed to address this issue. Both read and spontaneous speech were captured. The read speech segment comprises short stories chosen by the speaker for their literary style being easy to read out loud. The speaker familiarized herself with the content before reading them, in order to minimize speech errors and to produce prosody and articulation appropriate to the content. The spontaneous speech segment was produced by having the speaker describe some properties of images. Two examples include: (i) looking at a map and describing directions to travel between certain points, and (ii) looking at an grid of the "faces" method of multivariate data display and describing inferable data patterns.

3 COLLECTION PROCEDURES

Speech synthesis corpora demand much higher signal quality than suitable for other speech technologies. This requires minimization of background noise in the signal, phase distortion, and spectral distortion. Dual recording is highly desirable for accurate pitch extraction and identification of glottal closure. In addition, speaking style and consistency are two difficult and related issues which need to be addressed.

3.1 Noise Compensation

Standard professional recording procedures, such as high-end audio equipment and a structurally-isolated double-walled studio, were insufficient to control the background noise. We iteratively traced the sources of this noise by spectrally analyzing it, identifying the most prominent partials and resonances, and then isolating the relevant causes in the recording studio. We found, for example, that placing the pre-amplifier in the recording booth, and running it off batteries instead of grid power, reduced the 60 Hz harmonics and decreased the susceptibility to radio-frequency interference from computers in a nearby lab.

One persistent source of noise was the wideband hiss generated internally within the microphone itself. This was found to be true for a wide range of professional broadcast-quality microphones. We successfully reduced it by 2-3 dB by simultaneously recording from two such microphones next to each other and adding their (uncorrelated) signals. The two microphones were suspended in rubber shock absorbers to reduce vibrations from the floor, at a constant distance (5 inches) from the speaker's mouth.

Another source of noise in digital recordings is nonlinearities in the low-order bits of a digital-to-analog converter. To address that, we recorded direct to disk at 20 bits per sample and subsequently rounded to 16 bits. The SNR (computed spectral signal peak to spectral noise peak) was in the range 52-55 dB for all recordings.

3.2 Dual Recording

An electroglottograph (EGG) signal was recorded simultaneously with the acoustic signal. We used a single-electrode EGG device. In retrospect, this was largely but not completely successful. Subsequent analysis of the recordings showed that occasionally the speaker's larynx would move above or below the electrodes, and this would cause the signal to disappear momentarily. This occurred most often during the pitch peaks of nuclear accented syllables. We were able to minimize these dropouts by providing an oscilliscope display of the EGG signal in the speaker's line of sight. The speaker was thereby often able to monitor these dropouts, adjust the electrodes, and rerecord. Nevertheless this increased the cognitive load on the speaker and the overall duration of the collection. For this type of recordings, we recommend using multiple electrodes, or investigating sonar or radar techniques to capture laryngeal activity.

3.3 Speaking Style

Previous corpora have generally either left the speaking style up to the speaker, or have requested a professional news-reading style. Although the latter does tend to reduce variation in the signal amplitude, it implies a sustained level of vocal effort across all syllables of all words, which is not typical in normal conversation. (In regular day-to-day communicative speech, this degree of laryngeal tension and subglottal pressure is reserved for just the most salient syllables of the few most important words.) Since the spectral correlates are more high-frequency energy during voiced speech, we have found that the news-reading style results in a perceptually unpleasant, somewhat strident voice quality.

After exploring a few different styles we adopted a more relaxed, slightly more breathy speaking style. This style is used in the acting community to produce an impression of more intimacy. It is usually produced very close to the microphone, and is used, for example, when an actor is "talking to himself" or narrating his internal thoughts for the audience's benefit. We have found this to produce a softer, more pleasant voice quality in speech synthesis. Note that this relaxed laryngeal mode produces a larger open quotient in the larynx, and hence increases: (i) the proportion of subglottal coupling, and (ii) the proportion of nonharmonic noise (breathiness). Consequently the signal is less-well modeled by such frequency-domain representations as formant analysis or linear prediction. We use a time-domain signal representation, which is more robust to the corresponding assumption violations.

3.4 Consistency

A common problem in concatenative synthesis is that the units do not have consistent glottal slope, vocal effort, or perceived intensity. This entails discontinuities at the concatenation points which are not evident in the formant structure *per se*. It is difficult but crucially important to ensure that the speaker maintains a consistent speaking style, articulation rate, and vocal effort across the whole corpus. For reference, each session started with playing example recordings of the "model" voice quality, intonation, vocal effort, and pitch range. These examples were available all the time and often referred to by the speaker. For the majority of the recordings, an independent listener performed close, live monitoring of the intended production, correcting the speaker as necessary.

4 PROSODIC CONTEXTS

As mentioned earlier, this sub-corpus systematically models aspects of prosodic structure not well represented in the Polyphones sub-corpus. It focuses on:

(i) how pitch and duration of accented syllables varies with the distance to the next rightmost prosodic event [5], (ii) utterance-internal prosodic boundaries [6], (iii) final lengthening, and (iv) alignment of pitch and associated segmental structure in different syllable types. Each utterance consists of two accented words, preceded by several unstressed function words. The words were chosen to systematically vary the number of syllables between the two accented syllables between 0 and 5, and to systematically vary the position of the word boundary. Thus for example:

as a frill cheaply (2)

has the two accents (on "frill" and "cheap-") adjacent,

as a swamp customarily (3)

has two unaccented syllables between the accents (on "swamp" and "-mar-") with the word boundary adjacent to the left, and

the temerity throne (4)

also has two unaccented syllables between the accents (on "-mer-" and "throne") but this time with the word boundary adjacent to the right.

The design decision was made to only use real words (from the PRONLEX dictionary; cf. [1]), rather than to construct nonsense words to achieve the desired items. This was because we could not be confident that phrases of nonsense words would be spoken sufficiently naturally to represent normal English prosody. A grammar was constructed of English syllables, with phonemes collapsed into major classes. Words were chosen by a greedy algorithm to ensure that every possible syllable type occurred in each of the two relevant accent positions, and within each syllable type there was systematic variation of the instances of each of the phonemes in each of the classes. Thus, for example, in the syllable class "voiceless fricative, sonorant, high vowel, sonorant," if the word "frill" were used for one item, then the next time that syllable class was required a different word such as "swill" would be used instead.

The list generation procedure was completely automated to conform to these interacting sets of constraints, so that each list of phrases generated depended only on an initial condition (the first word chosen). Obviously, not all possibilities could be filled: sometimes there was no word available for a particular set of constraints. By using different initial conditions, different instances of the list of phrases were obtained. We recorded two different such lists. One contained 733 phrases, the other 631. Together these yielded 50,797 phonemes.

Each utterance was spoken in two ways: (i) a $\mathbf{H}^*$ pitch accent on both words (usually the accent on the second word was downstepped or reflected final lowering), and (ii) a $\mathbf{H}^*$ on each word, but with an intervening **L** tone associated with an utterance-internal intermediate phrase boundary.

The Prosodic Contexts sub-corpus is being used primarily for duration modeling. After collection, phoneme boundaries were automatically aligned using a speaker-dependent version of the Apple large vocabulary continuous speech recognition system. The resulting duration data was used to train the MacinTalk 4 statistical duration models described in [7].

5 CONCLUSION

The Victoria corpus collection effort has underscored a number of interesting lessons. First, despite superficial similarity, different criteria should be used for the design of speech synthesis and speech recognition corpora. In the latter, background noise and variability are desired, and lack of coverage of rare cases is often of little consequence. In the former, distortions, noise, and variability that characterize most real-world conditions will result in poorer synthesis quality, and coverage is necessary because most combinations of phonetic and prosodic contexts are rare. Second, collecting such a large speech synthesis corpus presents very real challenges. Consistency of vocal effort, pitch range, voice quality, and speaking rate across multiple months of recording sessions is as crucial as it is difficult. It helped to have spoken examples of the desired speaking style and intonational tunes available, and to make frequent reference to them.

6 ACKNOWLEDGEMENTS

We are indebted to D. Coombs for his audio engineering expertise and technical advice. S. Meredith and J. Butzberger generated the list of common content words and function word sequences, respectively.

REFERENCES

[1] http://www.ldc.upenn.edu/Catalog/index.html

[2] K.E.A. Silverman *et al.*, "TOBI: A Standard for Labelling English Prosody," in *Proc. ICSLP*, Banff, Canada, 1992.

[3] F. Kubala *et al.*, "The Hub and Spoke Paradigm for CSR Evaluation", in *Proc.* ARPA *Speech and Natural Language Workshop*, Morgan Kaufmann Publishers, pp. 40–44, March 1994.

[4] K.E.A. Silverman, *The Structure and Processing of Fundamental Frequency Contours*, Ph.D. Thesis, Cambridge University, 1987.

[5] K.E.A. Silverman and J.B. Pierrehumbert, "The Timing of Prenuclear High Accents in English," *J. Acoust. Soc. Am.*, Vol. 82, Supp. 1, 1987.

[6] K.E.A. Silverman, "Utterance–Internal Prosodic Boundaries," in *Proc. Austral. Conf. Speech Science Tech.*, 1988.

[7] J.R. Bellegarda and K.E.A. Silverman, "Improved Duration Modeling of English Phonemes Using a Root Sinusoidal Transformation," in *Proc. ICSLP*, Sydney, Australia, 1998.