SPEECH SYNTHESIS

For most people, speech has always been the most natural and preferred means of communication. Now speech with machines is required to provide a friendly way of accessing online information and enabling portable computing by using simple devices. Although, speech synthesis or computer-generated speech has been with us for more than half a century, however, it is still a relatively unused technology.

Speech synthesis is of particular interest as a hybrid science that involves interaction between natural language processing and digital signal processing, including such varied disciplines as linguistics, phonetics, phonology, electrical engineering, signal processing, computer science, mathematics, psychology, statistical modeling, speech science, and psychoacoustics. It is a branch of artificial intelligence that attempts to model one of the most subtle aspects of human performance.

One of the reasons that speech synthesis has not yet gained general acceptance in the marketplace is that although intelligible, the quality has usually been inadequate to portray the fine distinctions of meaning to which humans have grown accustomed. Speech synthesizers were designed as reading machines but have not been endowed with the intelligence to understand the words they speak.

There is a fundamental difference between speech (what humans do to communicate ideas) and text (the two-dimensional documentation of those ideas). Speech is not simply spoken text, nor is text just written speech. Humans learn to speak early and with very little instruction, but their literary skills take much more time and effort to acquire.

Speech is a one-dimensional time-dependent process that involves a speaker and a hearer in a situation that usually requires them to be simultaneously present, albeit remotely, such as over a telephone line. Text is relatively timeless. It is usually carefully composed according to a precise set of rules and is explicit and concise. Translating from one medium to the other is often difficult for humans and requires extensive background knowledge about the state of the world, the transmitting and the receiving agents (writer/speaker and reader/hearer), and the intentions underlying the communicative act. Because very little of this knowledge is available to machines required to perform the transformation automatically, the "speech" output from a synthesizer is usually only a poor copy of what a reasonable human expects.

MOBILE COMPUTING IN THE INFORMATION AGE

Two major recent developments in computing have had serious implications for speech synthesis technology. One is a rapid increase in use of the Internet, resulting in a proliferation of freely available information worldwide, and the other is a similarly rapid increase in available memory capacity and computing power which, combined with the development of small notebook computers and telephony interfaces, has enabled portable or mobile computing. When people on the move need to access information, speech becomes the medium of choice.

Speaking Machines

People on the move do not generally have the time or the patience to listen to long passages read aloud. Instead, they need access to limited chunks of information such as the location of the next turnoff, the latest news headline, or the subjectline of an email message. They need to be able to interact with the information provider, through speech, to obtain further related information, such as the distance to the turnoff or the name of the message sender.

Until recently, speech synthesizers were designed as reading machines. With the growth in mobile computing, however, they are now required to be speaking machines instead, conveying information to people whose eyes may be occupied elsewhere. There is a bigger difference in the requirements of the two types of application than the small difference in name implies.

Text and Speech

Most people absorb and retain a greater amount of information in less time from a page or a screen than from the one-dimensional medium of speech but, as noted previously, the majority of written texts are designed primarily to be looked at and not to be read aloud. As a consequence, literary or written sentences are longer than spoken ones and are generally more elaborate in their construction and use of language.

The density of information on a written page is typically much greater than that carried by the same number of words in a spoken utterance, and much important structural information is gained from the layout of a written text. This information can be lost when the text is simply converted into speech. Think, for example, of tables, figures, equations, and the use of different font types, sizes, and layout on the printed page.

Nonspeech Sounds

Human spoken interaction, on the other hand, makes much use of suprasegmental information and extralinguistic, nonspeech sounds to assist short-term memory for speech comprehension and to convey fine interpretations of meaning and speaker intention.

In face-to-face conversation, eyecontact, gesture, and facial expression are also used in conveying information, but the fact that most people have no difficulty talking to each other by telephone (and that many people actually prefer using a phone) shows that this extra information is not necessary for spoken discourse.

Nonspeech sounds, however, are normally used in conversational speech even in remote discourse, and often such sounds as breaths, sighs, pauses, and even sniffs and clucking of the tongue are used communicatively. We often laugh when we speak, and we add sounds to our speech to express emotion and feelings and to portray meaning.

If the next generation of speech synthesizers are to assist in human communication, then they should be capable of interaction with people using similar short and friendly "spoken" language, but this requirement brings with it the need for a new type of text analysis.

Interactive Information Access

Text-to-speech synthesis for spoken interaction with computers needs to be intelligent. For example, in Internet-based or in-car information access, the source text may often be structured and ideally suited to short targeted questions.

The visually-oriented structuring of World Wide Web documents are not easily converted into speech, yet they are now probably the most common written medium for computer-based public information access. However, it is rare that a web page can be simply read from top to bottom, and even the definition of "bottom" is no longer clear in a hypertext markup language (HTML) document in which many intermediate links take the reader to different sections or different documents.

Furthermore, neither the information provider nor the synthesizer can be expected to know what information the user might want or in what order to present it best at any given time. Therefore retrieval of clearly defined type-limited information through speech is better suited to question and answer interaction rather than passive listening.

Because people use extralinguistic information so much to aid in interpretating their words in natural spoken interaction, it is reasonable to expect it to be used equivalently in computer speech as well. However, the generation of intonation and nonspeech noises is not yet well enough developed in mechanical synthesizers, although recent advances in large-corpus concatenative systems may eventually enable the duplication of all speech noises.

TEXT PROCESSING FOR SPEECH SYNTHESIS

In the following sections we examine some of the processes required for converting linear text into speech and consider some of the different ways that machines produce speech. These sections provide no more than a set of pointers to highlight various aspects of the technology. They will help form a basis of understanding to give the general reader a more informed access to the literature and to provide the specialist with a brief overview of the latest developments in speech synthesis up to the turn of the century.

Figure 1 shows a flow diagram for a typical text-to-speech synthesizer. Text presented as input is first processed by lexical, morphological, and syntactic analysis before being converted into a sequence of phonemes. After this analysis, prosodic processing of the word sequence gives a specification of the appropriate pitch, power, and duration values for each phone in turn, and then a speech waveform is created according to these specifications.

The degree of information produced by the morphological and syntactic analyzers is usually quite limited because of the difficulty of resolving the many ambiguities common in a written text without access to world knowledge or discourse context and history. Similarly, because the precise meaning of an utterance to be spoken cannot usually be known to the synthesizer without special markup of the input text, only a simple default specification of the required prosody can be generated.

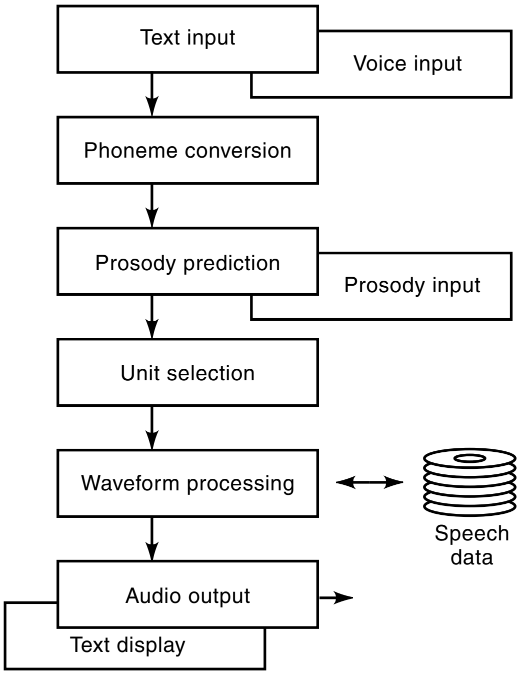


Figure 1. Flow of processing in text-to-speech synthesis. Input is normally plain written text, but may be structured text, annotated text, or even voice, for example, when a particular speaking style of intonation is required from a given synthetic voice. Input is mapped into phonemes, their prosody determined, and units for waveform generation are selected. Although output is usually in the form of voice, display of the intermediate results can be in the form of text output.

The following order of processing is typical: text preprocessing, lemmatization of word forms, accent assignment, word pronunciation, intonational phrasing, phrasal accent assignment, segmental duration prediction, fundamental frequency, F_0, prediction, amplitude assignment, waveform source-parameter computation, and concatenation of waveform units.

Text Input

Text to be spoken by a synthesizer must first be converted into a stream of phones or symbols that represent the individual sounds of the speech. To do this for a language like English, the spelling first needs to be disambiguated so that similarly spelled words like "record" (noun) and "record" (verb) can be differentiated. For this task, some morphological and syntactic analysis must be performed.

In many languages, there is a one-to-one mapping between the spelling (or written form) of a word and its pronunciation, but this is not always the case. In English, for example, although only five letters represent the vowels, there are at least fifteen different vowel sounds. The difference between the pronunciations of such similarly spelled words as "through", "thought", "cough", "bough", and "though" illustrates the diversity of the sounds of the language and the difficulty of predicting the sound from its spelling alone.

Dictionaries

The first requirement for spelling-to-sound processing is a dictionary. Because the number of words in a language is close to infinite, however, and because no single dictionary can be guaranteed to contain all the words that might be present in any given text to be synthesized, a set of letter-to-sound conversion rules is also required. Furthermore, because of the size constraints of many synthesis applications, the dictionaries used should be as compact as possible.

Dictionaries for text-to-speech synthesis need contain only the root forms or base forms for the pronunciation of words that cannot easily be predicted by the letter-to-sound rules. Morphological decomposition is performed to derive the base form from the lexical realization present in the text. For example, the word “flies” can be represented as consisting of the base form of the verb “fly” modified by the third person singular marker, subject to a rule of the form “y → ies”. It is not necessary to have separate dictionary entries for all of the derived forms (flying, fliers, flew, flown, flights, etc) if they can be reduced by a set of rules to a simple base form plus modifier(s). However, overgeneralization of such decomposition rules can be dangerous. For example, they might fail to analyze the orthographically similar word “lies”, because in English there is no equivalent underlying base form “ly”. The science of morphological decomposition has its edges in art.

Morphological Decomposition

One particular use of the information derived from such morphological decomposition, or word analysis, is for estimating the syntactic class or part-of-speech of the words in the text. For example, it may be very important for the prediction of an appropriate prosody to know that a given word “flies” is a verb, and not a plural noun. The often used example sentence “Time flies like an arrow” has a very different intonation depending on whether the first word is parsed as an adjective, a verb, or a noun. Suffixes like -er, -ing, -ed, etc., can be very informative for deriving the syntactic class of an orthographic word, but position in the utterance and classes of neighboring words are equally useful sources of information.

Once the syntactic class of a given word in an utterance has been estimated, its pronunciation and prosody can be determined (to a large extent) from position- and context-sensitive rules. Figures 2 and 3 show some of the rules for determining the pronunciation of a given orthographical sequence if it is not found in the dictionary. For a large number of words (such as “record” in the previous example), however, the pronunciation can be decided only by dictionary lookup in conjunction with syntactic part-of-speech information.

Unfortunately, there is no guarantee that a machine can successfully parse a given input text and in many cases default word classes have to be assigned. Because of this inherent uncertainty in text preprocessing, many synthesizers make only limited prosodic predictions on the principle that underspecification is a lesser mistake than an incorrect interpretation.

Expansion of Abbreviations

Not all text is made up of words, however, and abbreviations and numbers can provide particular problems for a text analyzer. Whereas simple abbreviations such as “%” for “percent” can usually be easily converted into the corresponding words, many are ambiguous: Dr. Smith Dr. or St. John St. would probably be read as “Doctor Smith Drive” and “Saint-John Street” respectively, with the same abbreviations expanded in two different ways each time.

```
D_rules[] = {
        {“#:”, “DED”, Nothing, “dIHd”},
        {“.E”, “D”, Nothing, “d”},
        {“#^:E”, “D”, Nothing, “t”},
        {Nothing, “DE”, “^#”, “dIH”},
        {Nothing, “DO”, Nothing, “dUW”},
        {Nothing, “DOES”, Anything, “dAHz”},
        {Nothing, “DOING”, Anything, “dUWIHNG”},
        {Nothing, “DOW”, Anything, “dAW”},
        {Anything, “DU”, “A”, “jUW”},
        {Anything, “D”, Anything, “d”}};
Y_rules[] = {
        {Anything, “YOUNG”, Anything, “yAHNG”},
        {Nothing, “YOU”, Anything, “yUW”},
        {Nothing, “YES”, Anything, “yEHs”},
        {Nothing, “Y”, Anything, “y”},
        {“#^:”, “Y”, Nothing, “IY”},
        {“#^:”, “Y”, “I”, “IY”},
        {“ :”, “Y”, Nothing, “AY”},
        {“ :”, “Y”, “#”, “AY”},
        {“ :”, “Y”, “^+:#”, “IH”},
        {“ :”, “Y”, “^#”, “AY”},
        {Anything, “Y”, Anything, “IH”}};
```

Figure 2. Converting orthography into phonemes. Rules are made up of four parts: the left context, the text to match, the right context, and the phonemes to substitute for the matched text. First, separate each block of letters (apostrophes included), and add a space on each side. Then for each unmatched letter in the word, look through the rules where the text to match starts with the letter in the word. If the text to match is found and the right and left context patterns also match, output the phonemes for that rule, and skip to the next unmatched letter. (Derived from Automatic Translation of English Text to Phonetics by Means of Letter-to-Sound Rules which was released into the public domain as NRL Report 7948 on January 21st, 1976 by the Naval Research Laboratory, Washington, DC.)

Numbers must be parsed for their interpretation before they can be pronounced. Numbers in an address may not be pronounced in the same way as the equivalent numbers in a count: we say “one-oh-one Park Lane”, but “a hundred and one Dalmations”, and four-digit number strings can provide a

IY	bEEt	IH	bIt	EY	gAte
EH	gEt	AE	fAt	AA	fAther
AO	lAWn	OW	lOne	UH	fUll
UW	fOOl	ER	mURdER	AX	About
AH	bUt	AY	hIde	AW	hOW
OY	tOY	U	YOU		

p	Pack	b	Back	t	Time
d	Dime	k	Coat	g	Goat
f	Fault	v	Vault	TH	eTHer
DH	eiTHer	s	Sue	z	Zoo
SH	leaSH	ZH	leiSure	HH	How
m	suM	n	suN	NG	suNG
l	Laugh	w	Wear	y	Young
r	Rate	CH	CHar	j	Jar
WH	WHere				

Figure 3. Machine-readable American English phonemic notation used in the orthography-to-phoneme rules shown in the previous figure (see Fig. 2).

special problem: “1950” could represent years (nineteen fifty), a count with the comma missing (one-thousand, nine-hundred and fifty), or a phone number (one, nine, five, zero).

Formulas and equations also have special pronunciation rules. The difference between “(1 + 5) × 9” and “1 + (5 × 9)” is easily seen on the printed page but requires particular phrasing to be realized in speech.

Inserting Pauses

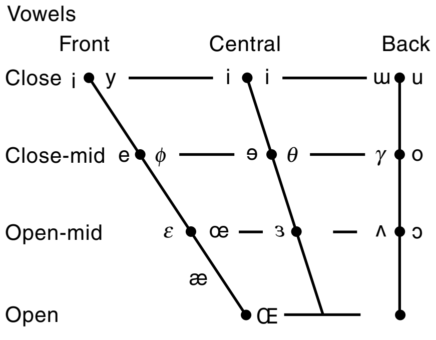
In much the same way as a piece of music requires specification of rests in the time framework for its ultimate realization, converting text into speech also requires predicting the number and length of any pauses between the words.

A numerical pause potential can be assigned to the boundary of every pair of words in an utterance, for example, by using a grammatical-category transition matrix to assign low potentials to commonly occurring progressions, such as subject-verb-object-modifier; slightly higher pause potentials between long subject and verb, and between object and trailing modifier; and relatively high potentials for reversals in the common sequence of grammatical categories, depending on the specific category transition. Alternative methods use length of constituent as a criterion, inserting a pause to balance each part of the sentence by reducing longer phrases into a sequence of shorter ones according to a weighting measure.

The realization of these potentials into actual pauses having specific duration further depends on such factors as the rate of speech, length of utterance, and various speaker-specific settings. Furthermore, “pauses without duration” can be realized for faster speaking styles by such devices as elongation of the prepausal vowel, a downward glide of the pitch contour just before the pause point, and local resetting of the pitch contour after it.

Basic Speech Sounds

The International Phonetic Association (IPA) has produced a description of all of the sounds (or “phones”) of the world’s languages, and has prescribed a standard written form for each. Figures 4 and 5 illustrate some of these sound-symbol mappings and their relationship to the speech production process. Although the commonly used IPA symbols are not easily computer-readable, most synthesizers use a machine-readable ASCII version or an equivalent sound-to-symbol mapping set to specify the individual speech sounds.



Where symbols appear in pairs, the one to the right represents a rounded vowel.

Figure 4. The International Phonetic Alphabet—Vowels—This chart shows the cardinal vowels and their relationship to each other in terms of articulative position with respect to jaw opening on the vertical axis and vocal tract configuration on the horizontal axis. (Chart by courtesy of the International Phonetic Association.)

The symbol-level output of a letter-to-sound module often derived by simple lookup of the citation-form pronunciation of a word, however, is not always sufficient to predict how a given word should be pronounced in a particular context. Coarticulation of a speech sound with its previous and following neighboring sounds strongly influences the way each is realized in any given context or speaking style.

Natural Speech

These coarticulation effects, sometimes mistakenly called “sloppy-speech phenomena,” are actually helpful for correct interpretation of the meaning and register of an utterance. For example, the word “going” is pronounced very differently when it is a main lexical verb (as in the utterance “I’m going to France”) and when it functions as a grammatical or future tense marker (as in “We’re going to get married in June”). In the latter case, and especially in informal speech, “going to” may be reduced to “g’nta” or “gonna”, reflecting its less important semantic role in the sentence and aiding to comprehend the utterance as a whole by helping to focus attention on the main verb. In the latter example, a full pronunciation of the word “going” would probably give the sentence a feeling of particular emphasis or contradiction. Some of these effects have been formalized in the English language and are notationally signified by an apostrophe, as in the previous example of “I’m”. The common words “am”, “have”, and “will” are rarely pronounced fully in fluent speech and are frequently contracted to the apostrophized form in writing.

Similar coarticulation effects are also found at a lower level within the speech signal, such that individual sounds categorized under the same phone label may be produced very differently according to their phonemic and prosodic contexts. In English, for example, a plosive sound like /p/, /t/, or /k/ may have different degrees of aspiration depending on whether or not it occurs in a stressed syllable (e.g., “apple” and “apply”), or whether it is followed by a word boundary (“plate rack” and “play track”) or by a front rather than a back vowel (“keel” vs “cool”). Sounds like /m/ and /l/ are strongly influenced by their positions in the syllable (e.g., “mum” and “lal”), and have light and dark variants that are realized differently depending on whether they occur before or after the vowel.

When converting from a symbolic representation of the sound sequence of an utterance to a lower level representation, such as that used to generate the speech signal, coarticulatory interactions play an important part. However, the degree of importance of predicting such coarticulation effects varies depending on the method used for speech signal generation. We return to this point in a later section, after first examining how the prosodic characteristics of each sound are determined.

PROSODIC PROCESSING FOR SPEECH SYNTHESIS

The first prosodic characteristic that must be determined is the segmental duration for each sound in the utterance. Un-

THE INTERNATIONAL PHONETIC ALPHABET (revised to 1993)

CONSONANTS (PULMONIC)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive			t d					k g	q g	□ ▨	□ ▨
Nasal			n							▨	▨
Trill			r					▨			▨
Tap or flap								▨			▨
Fricative	Φ *β*	f v	*θ*	s z	ʃ 3	f v					
Lateral fricative	▨	▨								▨	▨
Approximant			ɹ								▨
Lateral approximant	▨	▨	l							▨	▨

Figure 5. The International Phonetic Alphabet (Consonants). This chart shows the distribution of consonantal sounds in terms of place and articulative manner. The symbols are internationally standardized and unicode-supported. (By courtesy of the International Phonetic Association, c/o Department of Linguistics, University of Victoria, Victoria, British Columbia, Canada.)

like notes in a melody, which may have no inherent characteristics apart from their tonal difference, the durations of speech sounds differ characteristically according to where and how they are made in the mouth. For example, the vowel /a/ is typically longer in duration than the vowel /u/ because it requires more jaw opening. Similarly, the two consonant sounds /s/ and /r/ (as in "side" and "ride") are made at approximately the same place in the mouth (at the region joining the soft and hard palate) and with the same degree of jaw opening. But the /s/ is typically much longer in duration than the "r" because the tip of tongue has to be carefully placed close to the roof of the mouth and held there so that a special airflow is set up to produce the required sibilant energy. The /r/ sound is made by the simpler gesture of raising the tip of tongue toward the roof of the mouth.

The durations of the individual speech sounds also differ according to their phonemic and prosodic environments. An /i/ sound before a /t/ is likely to be shorter than a similar /i/ before a /d/ (e.g., "bit" verses "bid"), even though the /t/ and /d/ are made at the same place in the mouth and differ only in their manner of voicing. A sound at the end of an utterance before a pause is likely to be longer than a similar sound at the beginning or middle of an utterance, and the duration of a sound also varies greatly depending on whether or not it carries lexical stress (e.g., "man" vs. "fireman").

Rhythm

Knowledge of the stress patterns in a word is essential for predicting the duration of its segments, but articulatory stress is not governed by lexical defaults alone. The word "afternoon" spoken in isolation has three syllables. Lexical stress is normally carried by the third syllable, and only secondary or weaker stress on the first. However, in the context "afternoon tea" the stress moves back so that the primary stress is realized on the first syllable to prevent a "stress-clash" with the lexical stress carried by the monosyllabic word "tea". Such rhythmic rules need to be applied after the letter-to-sound rules but before duration prediction.

Figure 6 shows a set of rules for predicting the duration of each sound in its given context. These rules were derived heuristically from visual analysis of speech data, but more recent advances in synthesis research have resulted in automatic optimization of rules based on statistical analysis of large speech corpora. Figure 7 shows a tree-based version, where the rules were derived automatically and also shows the values predicted by each. The automatic learning of speech data characteristics is addressed separately in a later section.

Coarticulation

The articulatory characteristics of the segment itself are of primary importance, though how best to describe this factor is still a matter for active research. Although dark and light variants of the same phone, for example, have different articulatory and durational characteristics, they are often represented by the same phone label. Indeed, whether the notion of phones is valid for describing speech sounds at all is still very much open to question because it emphasizes the idea of independent entities strung together sequentially, whereas in

The value of P(%) is initially set to 100 then modified by each applicable rule $P = P \times \frac{P1}{100}$.

1. PAUSE INSERTION: Insert a brief pause (200 msec) before each sentence-internal main clause and at other boundaries delimited by an orthographic comma.
2. CLAUSE-FINAL LENGTHENING: (P1 = 140) The vowel or syllabic consonant just before a pause is lengthened. Any consonants in the rhyme (between this vowel and the pause) are also lengthened.
3. PHRASE-FINAL LENGTHENING: (P1 = 140) Syllabic segments (vowels or syllabic consonants) are lengthened in a phrase-final syllable. Durational increases at the noun/verb-phrase boundary are more likely in a complex noun phrase or when subject-verb order is violated; durational changes are much more likely for pronouns.
4. NON-WORD-FINAL SHORTENING: (P1 = 85) Syllabic segments are shortened slightly if not in a word-final syllable.
5. POLYSYLLABIC SHORTENING: (P1 = 80) Syllabic segments in a polysyllabic word are shortened.
6. NON-INITIAL CONSONANT SHORTENING: (P1 = 85) Non-word-initial consonants are shortened.
7. UNSTRESSED SHORTENING: Unstressed segments are shorter and considered more compressible than stressed segments. The minimum durations for unstressed segments are halved (MINDUR = MINDUR/2) then stressed and secondary-stressed segments are shortened: Consonants before a stressed vowel that are in the same morpheme or form an acceptable word-initial cluster are also considered to be stressed. (syllabic (word-medial syll): P1 = 50, syllabic (others): P1 = 70, prevocalic liquid or glide: P1 = 10, others: P1 = 70).
8. LENGTHENING FOR EMPHASIS: (P1 = 140) An emphasised vowel is significantly lengthened. This lengthening can also be used to capture word frequency and discourse effects that are not otherwise incorporated in the rule system.
9. POSTVOCALIC CONTEXT OF VOWELS: The influence of a post-vocalic consonant (in the same word) on the duration of the vowel is such as to shorten the vowel if the consonant is voiceless. The effects are greatest at phrase and clause boundaries (open syllable, stressed, word-final: P1 = 120, before a voiced fricative: P1 = 160, before a voiced plosive: P1 = 120, before an unstressed nasal: P1 = 85, before a voiceless plosive: P1 = 70, before all others: P1 = 100).
10. SHORTENING IN CLUSTERS: Segments are shortened in consonant-consonant sequences (disregarding word boundaries, but not across phrase boundaries) (vowel followed by vowel: P1 = 120, vowel preceded by vowel: P1 = 70, consonant surrounded by consonants: P1 = 50, consonant preceded by a consonant: P1 = 70, consonant followed by a consonant: P1 = 70).
11. LENGTHENING DUE TO PLOSIVE ASPIRATION: A stressed vowel or sonorant preceded by a voiceless plosive is lengthened. In contrast to all other modifications, which effect a percentage change to part of the segment's inherent duration, this is an additive modification by a fixed value of 25 msec.

Figure 6. The Klatt duration rules from D. H. Klatt. (Review of text-to-speech conversion for English, *J. Acoustic Society Amer.,* **82**: 737–793.)

actual articulation many sounds are simultaneously produced and may be better represented as combinations of features.

As an example, we can consider the word "sprint". In the orthography it is clear that the "p" comes after the "s" and before the "r", but in speech it is likely that the three sounds are articulated almost simultaneously and that they might be better represented as overlapping. By describing the speech events in a declarative, tiered representation using such features as nasality, labiality, laterality, closure, voicing, and aspiration, etc., we are able to model this coarticulation well and to predict better the elisions that occur in fast or fluent speech.

Figure 8 shows how the word "pleasure" is represented by declarative constraints in a nonphonemic way, and Fig. 9 illustrates a feature-based parametric representation of the articulatory processes that allows predicting speech in a nonphonemic way.

Syllables as Basic Units

Many theories of speech science reject the phone as a basic unit of description, but for the majority of synthesizers it remains the common unit of specification, perhaps because of the ease with which it maps from words to waveform through an intermediate pronunciation dictionary.

Extending the notion of basic speech unit higher up the representational hierarchy, we find the syllable, the foot, the prosodic word, and the phrase also proposed as basic units. Several models of duration prediction use these higher level units as frameworks to constrain the limits of expansion or reduction of their component segments by applying a multilevel form of prediction to derive an overall time framework for the utterance.

There is considerable support for the notion of the syllable as a basic unit of articulation. In this view, speech is perceived as a sequence of vocalic sounds overlayed with contoid or consonantal modulations. The basic rhythms of the speech are governed by the strength of the syllables which in turn are decided by their compositional and contextual factors.

Syllable duration can be predicted externally from its position in the utterance with respect to semantic and discoursal features, such as information content and phrasing, and internally from the nature of its segments. Inherently long vowels and consonants have a lengthening effect and semantically important positions likewise. The overall duration is determined as a combination of these higher and lower level constraints.

If the syllable is a valid unit for describing speech, as many believe, then some doubt is cast on the optimality of the commonly-used phone-based description. Theoretical considerations aside, the use of the syllable as a basic unit to predict speech timing has the beneficial effect of limiting prediction errors and preserving rhythmical regularity, reflecting observations that an overlengthening of one segment is less likely to be perceived if it is accompanied by a corresponding shortening of a neighboring segment within the same syllable.

Contextual Influences

We have seen how segmental articulation has a bottom-up effect on syllable duration. Now we turn our attention to the top-down or higher level contextual effects. These are best described in terms of “salience” and “segmentation”. Salience is determined by pragmatic and discoursal forces, such as prominence or focus, information content, and relevance, whereas segmentation is a result of relationships between words and the grouping of words into phrases.

In determining salience factors for predicting speech timing effects, recourse is usually made to a buffer or stack of previously mentioned items in a text so that “given-ness” to the discourse can be estimated. At the simplest level, part-of-speech information can be used to estimate the contribution of a lexical item to the meaning of an utterance. Nouns and verbs are more salient than closed-class words, such as prepositions or determiners, resulting in longer durations for more salient items. The scope of this salience varies from a whole phrase to as little as a single phone (or less), so that in the example sentence, “I didn’t say *cake,* I said *take*”, the word “take” is more clearly articulated, resulting in a longer than normal duration for the closure of the /t/.

Segmentation is perhaps easier to identify because there is often a close correspondence between the syntactic phrasing of a text and the prosodic phrasing of an equivalent utterance, with the result that boundaries occur in similar positions. Whereas it might be a mistake to imply a causal relationship between syntax and prosody, they both function to delineate chunks by linking some parts of the structure more closely than others. There are many cases, however, where an identical part-of-speech sequence can have several different syntactic or prosodic bracketings depending on semantic factors. The common example “I saw the man on the hill with the telescope” requires an indication of the role of the telescope before it can be disambiguated in speech. If the telescope is used for viewing, then “the man on the hill” is uttered as a group with a short pause following. Otherwise “the hill with the telescope” with no pause might be a better interpretation.

Both salience and segmentation can be interpreted as scalar factors with different degrees of strength for each and with a positive correlation observed between the strength of the factor and the lengthening of the speech segments concerned. The nature of that lengthening depends on the type of the factor, such that salience has an effect better observed

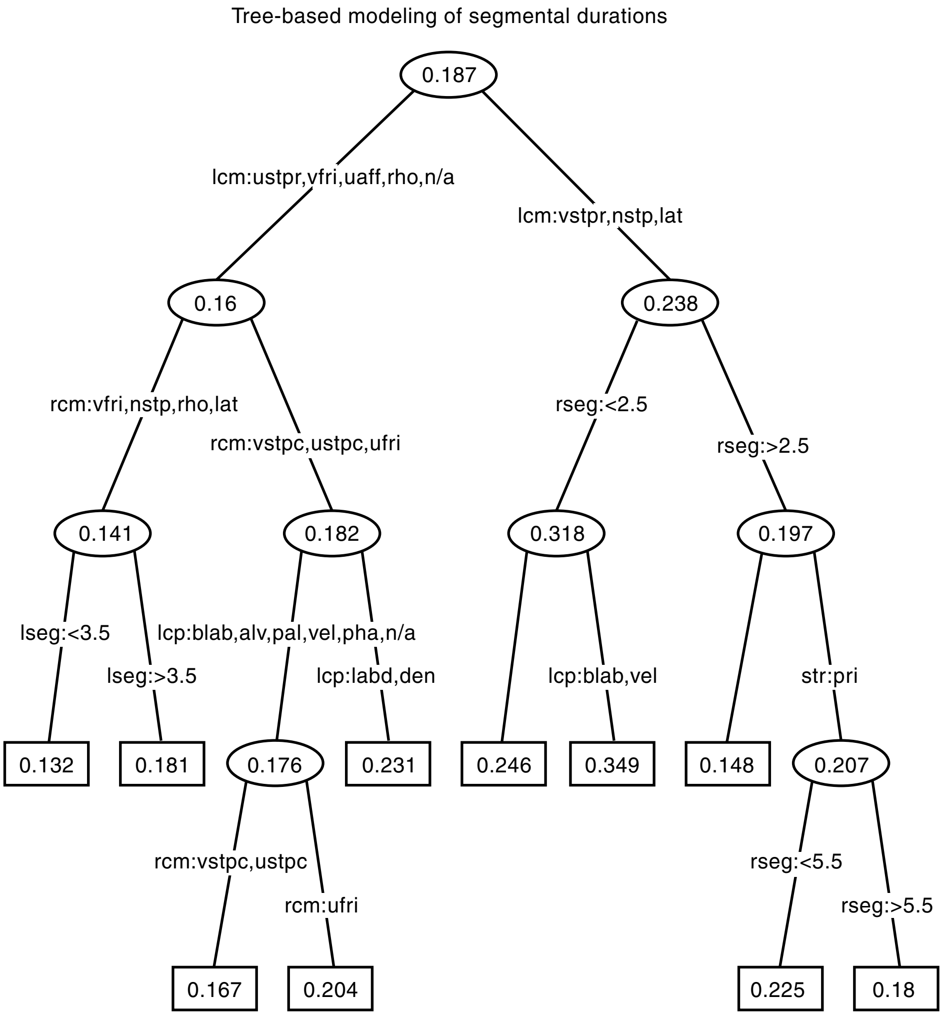


Figure 7. Part of a tree-based duration predictive rule system.

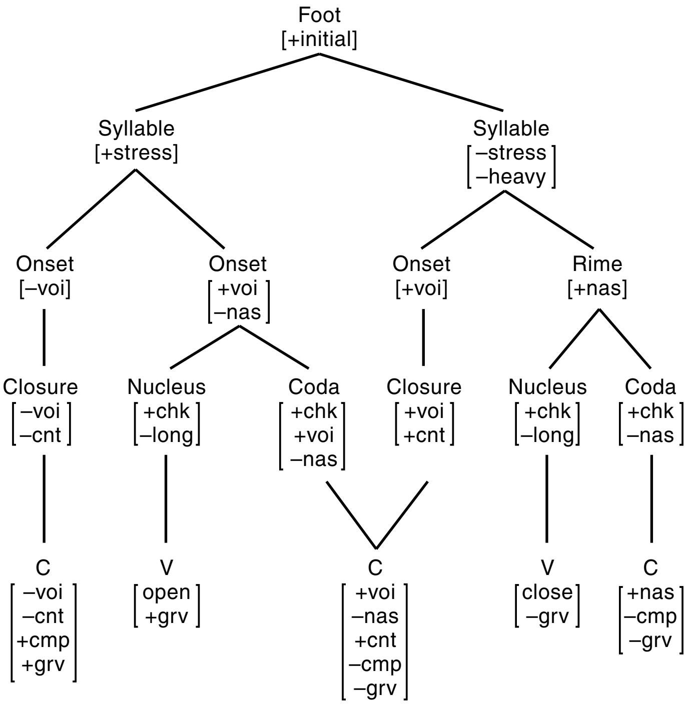


Figure 8. Part of a nonsegmental phonological representation of the word "pleasure."

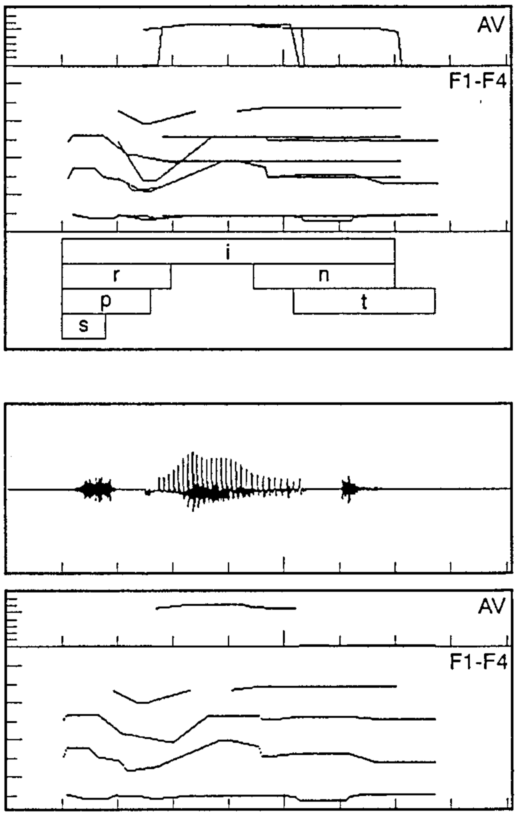


Figure 9. Phonetic interpretation of the word "sprint" and the parameter generation for its waveform.

on segments early in the syllable (in the onset and peak), and segmentation on those of the rhyme or coda. These form two sides of a virtual triangle of lengthening effects, and overall speaking rate forms its base.

Predicting Amplitude

Because the factors that control the durations in speech also have a related influence on the amplitude of the speech segments, similar models can be used to predict both the duration and the power of each speech signal segment. Stressed syllables have longer durations than their unstressed counterparts and also are correspondingly louder.

Although the controlling factors may be the same, however, segmental amplitude and duration do not always correlate positively. In utterance final positions, for example, we typically observe increases in duration but decreases in amplitude, as the speech slows down and decays into the following pause.

Pitch and Meaning

Another attribute that covaries with amplitude is the pitch of the voice as it changes over different parts of the utterance. Pitch is a subjective attribute, but its physical correlate in the speech signal, the fundamental frequency of vibrations of the glottal source, can be objectively measured and predicted. Once the duration of each segment in the speech stream has been predicted, a fundamental frequency contour can be determined for each component segment.

The information carried by variations in the fundamental frequency F_0 of the voice is probably as rich in meaning as that of the spectral variation in an utterance. As the spectrum defines the segments and differentiates one phone from another, so the F_0 signals the relationships between the words thus created, marking focus, delimiting phrases, and differentiating questions from statements. Therefore predicting an appropriate F_0 contour for synthesis is extremely complicated and ideally requires access to high-level information about the meaning and intentions underlying the content of the utterance.

Predicting the F_0 Contour

It is common to consider that the F_0 contour is made up of several component contours of varying scope. The main component, that of a single utterance, can be considered a carrier signal declining over time. The declination is reset at major prosodic boundaries and signals the internal coherence of each group of component phrases. Paragraph-level declination has been observed but is rarely modeled in current text-to-speech systems. A contrasting view has recently been advanced that this declination effect is more accurately represented as an edge-marking device, calling into question the nature of the decline in the mid-portions of longer utterances. But for short utterances there is agreement on the general effect.

Overlaid on the utterance-level carrier are phrase-level, syllable-level, and phone-level components. Figure 10 shows one such representation of the F_0. It is a simplification that does not account for any paragraph-level or phone-level effects, but it illustrates how a multicomponent contour can be conceived and realized. An alternative to this superpositional

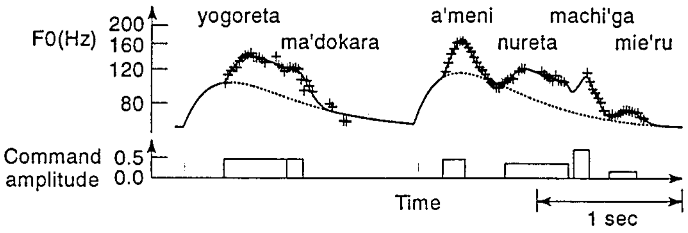


Figure 10. Superpositional modeling of F_0.

view is that of the tonal theorists, who model the contour as a simple sequence of high or low tones marking events related to the several layers of meaning in a speech signal. In the tonal view, there are only two basic tone types, high and low, and the contour is made up of tonal events at the word, phrase, and utterance levels. Figure 11 illustrates a speech waveform with its associated fundamental frequency contour is labeled according to the ToBI view of tonal sequences.

Normalizing Pitch Ranges

The speech fundamental frequency encodes information about the content of an utterance and also about its context, about who is speaking, and how. The average and range of voice fundamental frequency vary greatly between men, women, and children, and also according to such factors as emotion, speaking rate, and mood. For example, the typically low F_0 of a depressed speaker may not vary as much as that of a healthy speaker, and an angry speaker has a higher F_0 with less variation than that of an excited speaker. It is rare that such rich contextual information is available to a speech synthesizer from the raw text alone, but it may be essential to interpret a complete message successfully.

Methods of modeling F_0 can be made independent of such variation in speaker and range by using normalized values, such as *z*-scores, to convey information about the shape of a contour, which then can be rescaled to the desired range at a later stage of processing. The *z*-score transform is commonly used in psychometrics, but has recently been applied to prosodic prediction. It is computed by subtracting the mean (usually computed for each phone type individually) and dividing by the standard deviation of the distribution to express the data as a unitless number in the normally distributed range of plus or minus three.

Modeling the Contour

As with the modeling of segmental durations, there are many methods used to predict numerical values for F_0. All take as input some representation of the accentual patterning of the utterance, its phrasing, and a time sequence for the underlying phones. They differ primarily in how they model the contour, either as a superpositional hierarchy or as a series of shorter linear sequences, and in whether they are trained from corpus examples or derived from heuristic rules.

The corpus-based methods use large numbers of training examples derived from real speech in conjunction with statistical learning methods, such as neural networks, binary classification trees, or linear regression models. Typical input factors are the number of syllables, the stress and accent of each, and their parts of speech and syntactic bracketings. Outputs are either the direct numerical value representing one or several F_0 points for each syllable or a sequence of tonal representations (H or L for syllable and phrase) for subsequently predicting numerical values.

Aligning the predicted contour to the phone sequence is usually done by interpolation between the target points pre-

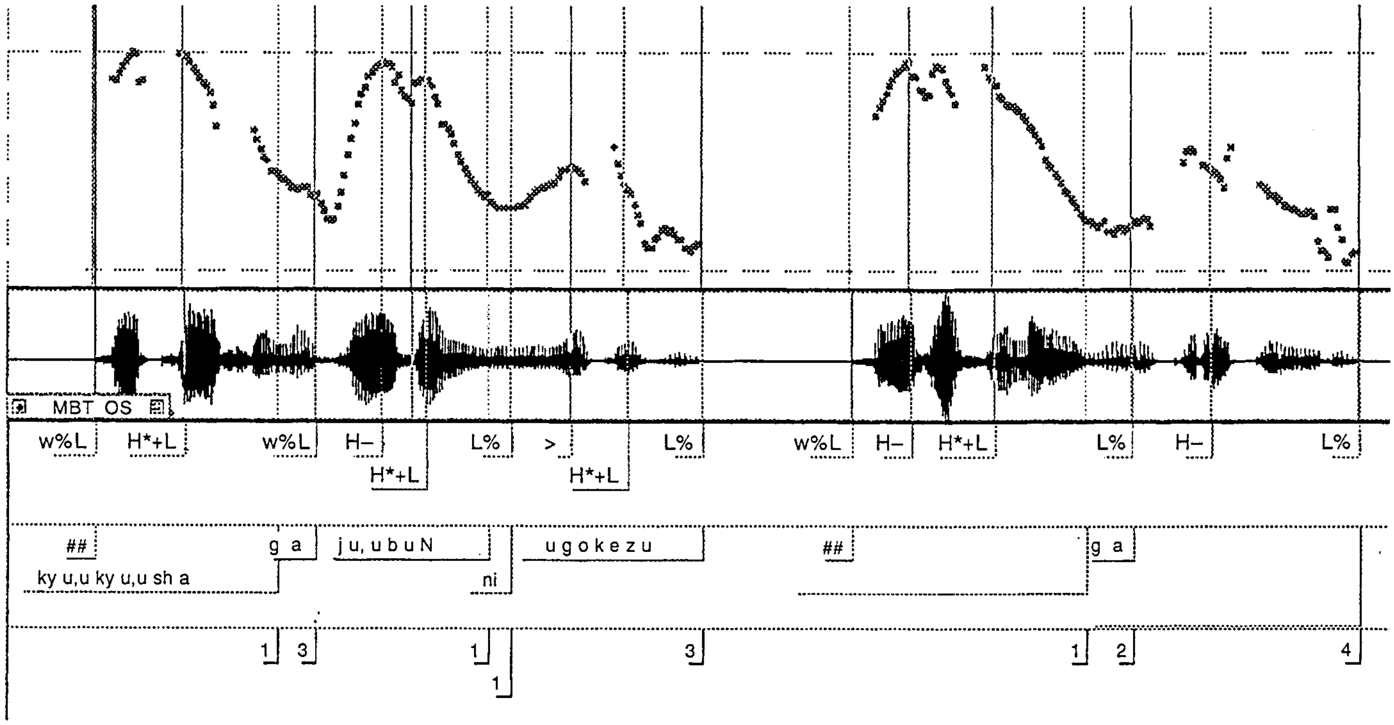


Figure 11. Tonal representation of F_0.

dicted for the key syllables, either by straight lines or by quadratic splines. It is generally accepted that a stylized contour is perceptually sufficient, and although in real speech there is considerable influence from the individual phonemic segments (nasals lowering the contour locally and plosives disrupting it characteristically), there has been little increase in perceived quality from modeling the microprosodic variations arising from the segmental nature of the utterance.

Depending on the synthesis method used, a degree of jitter or randomization is added to the interpolated contour to reduce the smoothness and to prevent an artificial "ringing" effect in the synthesized speech that can be caused by sustained level pitch on vowel sounds.

Once the segment stream and its prosodic characteristics have been specified, the higher level analysis and prediction phases of the speech synthesis are complete, and the system begins to implement the specification to create a speech waveform. This marks the crossover between the two main stages of synthesis, natural language processing (NLP) and digital signal processing (DSP).

SIGNAL PROCESSING FOR SPEECH SYNTHESIS

Several methods have been proposed for generating a waveform for speech synthesis. They are generally classified into either parametric or concatenative types. The former use knowledge about the frequency- and time-domain characteristics of the speech signal to recreate a waveform by rule, and the latter generate one by using prerecorded segments of real speech.

Parametric waveform generation techniques can be further divided into articulatory (top-down) and signal-based (bottom-up) variants. The former model the physical attributes of the human vocal tract to reproduce the acoustic environments that generate the speech sounds, and the latter use information about the formant structure of the individual phones to model their acoustic sequences and combinations.

Articulatory methods offer the best potential for synthesis because the parameters are limited by the same constraints that govern human speech, but the parameter vectors are difficult to define. Although articulatory synthesis generates individual sounds that are indistinguishable from human speech, it has yet to demonstrate contiguous sequences of vowels complete with consonantal transitions of good enough quality for real-time speech synthesis. The control functions necessary for the dynamic aspects of speech production are particularly difficult to model.

Although considerable research is being carried out into articulatory methods, in part because they offer the most insight into human speech production mechanisms, formant-based and concatenative methods have predominated in practical speech synthesis applications.

The Speech Waveform

Before examining the details of waveform production, first it is useful to consider the nature of the speech signal. Speech sounds are produced by vibration of the glottis or by obstructing to the passage of air through the vocal tract, resulting in a speech waveform that is made up of both periodic and aperiodic components. The categorization of speech sounds into vowels and consonants reflects this difference. Vowels, semivowels, and sonorant consonants exhibit a regular periodic structure, and obstruents correspond to less regular components in the signal.

Linear acoustic theory describes speech production in terms of a source and filter model. This model is made of a volume velocity source, which represents the glottal signal, a filter associated with the vocal tract, and a radiation component, which relates the volume velocity at the lips to the radiated pressure in the far acoustic field. This decomposition is acceptable for phonetics, which describes speech in analogous terms. "Phonation" equates to "source", and "articulation" is represented by the "filter." From the viewpoint of physics, however, this model is only an approximation, whose main advantage is simplicity. It is considered valid for frequencies below 4 kHz to 5 kHz, where the assumption of plane-wave propagation in the vocal tract is acceptable.

The Glottal Source

The earliest techniques of waveform generation for synthesis used banks of filters to recreate the formant structure. MI-Talk used both parallel and serial sets of filters excited by a periodic or noisy source.

The following model of the voice source is used in the Klatt synthesizer. Motivations for this model were pragmatic with formant synthesis applications in mind. It is a composite model which contain three components: a noise component, a periodic glottal waveform U_g^k, which is passed through a spectral tilt filter. The periodic component of the model is characterized by four parameters: the fundamental frequency f_0, the amplitude of voicing AV, the open quotient O_q, and the frequency of a spectral tilt filter TL. The periodic and aperiodic components are added. The equation of the model is

$$U_{\mathrm{g}}^k(t) = at^2 - bt^3$$

with

$$a = \frac{27AV}{4T_0O_{\mathrm{q}}^2}$$
$$b = \frac{27AV}{4T_0^2O_{\mathrm{q}}^3}$$

After some calculation one can show that the spectrum of $u_g^k(t)$ with $\nu = 2\pi/\omega$) given by

$$\tilde{U}_{\mathrm{g}}^k(\nu) = \frac{27jAV}{2O_{\mathrm{q}}(2\pi\nu)^2}\left[\frac{j\exp(-j2\pi\nu O_{\mathrm{q}}T_0)}{2} + \frac{1+2\exp(-j2\pi\nu O_{\mathrm{q}}T_0)}{2\pi\nu O_{\mathrm{q}}T_0} + 3j\frac{1-\exp(-j2\pi\nu O_{\mathrm{q}}T_0)}{(2\pi\nu O_{\mathrm{q}}T_0)^2}\right]$$

Source and Filter

The acoustic model can be written directly in terms of linear systems in the domain of signal processing in so far as the source/filter interaction can be neglected, although it may actually be possible to account for some interactive effects in the source/filter model.

For instance the effect of glottal leakage, or breathiness, is simulated by increasing the bandwidth of the first formant in the filter, together with modifying the source parameters. In its simplest form, the source filter model is written as

$$\begin{aligned} s(t) &= e(t) \times v(t) \times l(t) \\ S(\omega) &= |S(\omega)|e^{j\theta(\omega)} \\ &= E(\omega) \times V(\omega) \times L(\omega) \end{aligned}$$

where $s(t)$ is the speech signal, $v(t)$ is the vocal tract impulse response, $e(t)$ is the vocal excitation source, $l(t)$ is the impulse response of the sound radiation component, and $S(\omega)$, $V(\omega)$, $E(\omega)$, and $L(\omega)$ are the Fourier transforms of $s(t)$, $v(t)$, $e(t)$, and $l(t)$, respectively.

This equation suggests that spectral processing should be easier than time-domain processing. The source component $e(t)$, $E(\omega)$ is a compound signal, which can be represented with the sum of a quasi-periodic component (described by its fundamental frequency and its waveform) and a noise component:

$$\begin{aligned} s(t) &= [p(t) + r(t)] \times v(t) \times l(t) \\ &= \left[\sum_{i=-\infty}^{+\infty} \delta(t - it0) \times u_g(t) + r(t) \right] \times v(t) \times l(t) \\ S(\omega) &= [P(\omega) + R(\omega)] \times V(\omega) \times L(\omega) \\ &= \left\{ \left[\sum_{i=-\infty}^{+\infty} \delta(\omega - if0) \right] |U_{\mathrm{g}}(\omega)|e^{j\theta_{\mathrm{ug}}(\omega)} + |R(\omega)|e^{j\theta_{\mathrm{r}}(\omega)} \right\} \\ &\quad \times |V(\omega)|e^{j\theta_{\mathrm{v}}(\omega)} \times |L(\omega)|e^{j\theta_{\mathrm{t}}(\omega)} \end{aligned}$$

where $p(t)$ is the quasi-periodic component of excitation, $u_{\mathrm{g}}(t)$ is the glottal flow signal, t_0 is the fundamental period, $r(t)$ is the noise component of glottal excitation, δ is the Dirac distribution, $P(\omega)$, $R(\omega)$, and $U_{\mathrm{g}}(\omega)$, are the Fourier transforms of $p(t)$, $r(t)$, and $u_{\mathrm{g}}(t)$, respectively, and $f_0 = 1/t_0$ is the fundamental frequency of voicing.

As far as intraspeaker voice quality is concerned, the most important component is the source component, which is described by r, u_{g}, and f_0. Modifying this component changes voice quality but not voice personality, and modifying the filter component alters voice personality but preserves voice quality. This is only an approximation, however, and it is actually necessary to modify both components to achieve realistic modifications of either voice quality or voice personality.

Linear Prediction of Speech

Because the speech waveform can be considered a relatively slowly changing signal most of the time, it is well predicted by linear prediction techniques. A linear predictor uses observations of a speech signal to try and predict the next sample of the signal beyond those which it can observe, and the filter coefficients change perhaps every 20 ms. When the linear predictor is working well, there is little residual correlation between the error signal and the samples.

In the perfect case, filtering periodic pulse excitation with the inverse lattice filter yields intelligible speech, but in practice the speech thus produced sounds mechanical. Although we can model the filter characteristics of the vocal tract and the static waveshape of the glottal pulse very accurately, it is also necessary to control their dynamic variation to reproduce the period by period perturbations of source pulses and formant ripple that occur in natural speech.

Celp Coding

Because of the difficulty of reproducing voice dynamics by rule, many synthesis systems code them explicitly and store the results of inversely filtering the speech waveform as a separate excitation component. To reduce the memory requirements of this quality improvement, vector quantization is applied to the excitation patterns, and they are stored in the form of codebook entries, allowing only the identity of a particular entry to be transmitted and thereby considerably enabling compression of the required information. Codebook excited linear prediction has become widely preferred over traditional formant techniques for synthesis.

PSOLA Transforms

Although parametric methods offer easy manipulation of the duration, pitch, and power of the speech signal, they are lossy encodings and the resulting synthesis, although usually highly intelligible and easily recognizable as speech, rarely sounds close to the human original. The Pitch Synchronous Overlap & Add (PSOLA) algorithm was designed to independently modify a raw speech waveform with respect to F_0 and duration. It quickly became very popular because of its relatively high quality speech output.

In PSOLA manipulation, the speech waveform is first windowed pitch synchronously using a Hanning window to produce a set of pitch synchronous short time (ST) signals two pitch periods long overlapping by one pitch period with each neighbor. Pitch-synchronous labeling of the speech is required for this purpose. To achieve pitch variations, the ST signals are shifted against one another in time and then added again (overlap & add → OLA). Shifting them together results in higher pitch, and shifting them apart lowers the frequency. Duration is changed by doubling or leaving out certain ST segments before the final addition process. Inserting additional ST signals into the speech slows it down, and taking segments out leads to shorter durations.

An important part of the PSOLA algorithm is the mapping between the ST signals in the input stream and the ST signals in the output stream, which is controlled by the flow of modification factors. Careful study of the time synchronization leads to a mapping represented by the following equation:

$$t_{\mathrm{s}}(u+1) - t_{\mathrm{s}}(u) = \frac{1}{t'_{\mathrm{s}}(u+1) - t'_{s}(u)} \int_{t'_{\mathrm{s}}}^{t'_{\mathrm{s}}(u)} \frac{P(t)}{\beta(t)}\, dt \qquad (1)$$

where t_{s} denotes the pitch marks (and corresponding ST signals) of the input signal, t'_{s} are the pitch marks (and ST signals) of the output signal; $P(t)$ is the pitch period of the input signal, and $\beta(t)$ is the pitch scaling factor.

This integral equation is relatively easy to solve because the factors are piecewise linear. In a simple implementation, the procedure is as follows (assuming constant factors over a pitch period for explanatory purposes only):

1. At a specific instant, the output pitch period $P'(t)$ is determined by dividing the input pitch period at that time $P(t)$ by the modification factor $\beta(t)$.

2. Adding $P'(t)$ to the last pitch mark $t'_s(u)$ in the output stream gives us the next pitch mark $t'_s(u + 1)$ in the output stream.
3. Considering the duration modification factors up to this time (by integrating them), the point $t_s(u + 1)$ in the input stream corresponding to $t'_s(u + 1)$ is found.
4. The ST signal (i.e., pitch mark) lying closest to this point $t_s(u + 1)$ is mapped next. This is actually the mechanism described above: ST signals are doubled or skipped depending on the distance between $t_s(u)$ and $t_s(u + 1)$.

For factors changing within one pitch period, the algorithm becomes only slightly more complicated. In this case $P'(t)$ is also the result of an integration.

The PSOLA algorithm produces very natural-sounding speech for smaller modification factors. For good quality F_0 modification, factors should generally fall in the range between 0.7 and 1.3. The range for modifications of high-quality duration also depends largely on the nature of the phone to which the modification is applied. For longer vowels, stretching up to a factor of 2 still produces good results, but for shorter vowels (like schwas) even a stretch of 1.2 produces noticeable disruption to the speech.

Unfortunately, even signals having identical spectral envelopes and windowed on similar relative positions cannot be properly overlap-added if their periods are very different. So for a synthesis database it is necessary to find a reference speaker who is maintains a very even pitch, and this greatly limits the choice of voices available.

Furthermore, because it is a time-domain technique, PSOLA has no way of matching spectral envelopes of different concatenated segments together. So the speaker for a PSOLA concatenative synthesizer has to maintain a steady pitch and also has to be spectrally consistent. These characteristics do not produce lively and spontaneous-sounding speech.

Spectral-Domain Models

Harmonic decomposition of a speech signal was introduced to overcome PSOLA's limitations with respect to the range of modification that can be performed. This method treats the speech waveform as a summation of both harmonic and noise components at all frequency levels and by modeling the harmonics as a series of sinusoidal signals, produces an estimate of the residual or component noise at each frequency band.

The speech waveform is first inversely filtered to derive an approximation to the voice source. It is possible to directly modify the speech signal rather than the source, but in this case both source and filter are modified, and the effect of spectral modification does not correspond accurately to the previous equations.

Then the two components of the source signal are separated by periodic-aperiodic decomposion before periodic component mModification. The periodic component U_g is modified according to the previous equations cited. Zero-phase filtering implements modification of H_1, H_2 ($. . . H_n$), and of spectral tilt (any aspects of the amplitude spectra can be modified).

Finally, the aperiodic component is modified in the spectral domain with respect to amplitude and modulation, and Then the modified source signal is reconstructed by adding the modified periodic and the modified aperiodic component. At this stage it is also possible to modify the vocal tract parameters by adjusting the synthesis filter.

CONCATENATIVE SPEECH SYNTHESIS

Whereas it is relatively easy to model and manipulate the slowly changing characteristics of steady states of the speech sounds, such as found in clearly articulated or sustained vowels, the transitions from phone to phone between these representative target states in natural speech prove much more difficult to represent by rule or to modify. The articulation of each phone depends on its spectral and its prosodic contexts, and in fluent speech it is common that "steady states" are not reached at all.

Because of the difficulty of accurately modeling the intersegmental transitions in fluent speech, single-phone models of the speech signal proved difficult to work with, and diphones excised from naturally spoken utterances soon became accepted as the basic unit of synthesis. This prompted a reduction in rule-based waveform modeling and the marked the beginning of concatenative synthesis.

A variety of research followed into the nature of the recorded database, the size and type of units stored, the use of multiple units for different contexts, and different types of parametric representations.

Diphone Synthesis

A diphone encodes the transitions between a given pair of phones and, because of relatively steady or only slowly changing spectral characteristics at its edges, enables simple concatenation while encoding the more subtle interphone transition information internally. Diphones cut from the recorded speech waveforms incorporate a large amount of spectral information from the human speech signal for concatenative synthesis.

In English, the number of phones required to synthesize any word is approximately 45, depending on the dialect, and the equivalent number of diphones is typically a little over 2000, depending on the system. Not every phone-phone pair is realized in the language. But in most diphone-based synthesis systems, some interphone sequences are coded as triphones to model the 'weaker' transitions through phones like /h/ and schwa which are likely to be influenced simultaneously by the sounds on both of their edges.

Diphones encode the spectral transition information well and are easily concatenated because they are cut at the most stable point in each phone. But they do not encode the prosodic variation and so are usually stored in parametric form for subsequent modification before being concatenated as speech. Because the encoding, manipulation, and decoding stages of prosodic modification each incorporate a degree of degradation of the resulting waveform, the ensuing speech is often of limited naturalness and not always recognizable as that of the source speaker, even though it is highly intelligible.

Large-Corpus-Based Synthesis

To improve the recognizability and naturalness of concatenated speech and to reduce the degradation resulting from signal processing, several large-corpus-based methods have recently been proposed which extend the trend to increase unit size and number at the expense of storage capacity.

To account for as many contextual effects as possible, context-oriented clustering has been proposed for automatically determining an optimal set of speech units to be cut from natural speech. This tree-based, data-clustering method starts with a database labeled at the phone level and continues to split clusters of phones with the highest acoustic variance until a flat distribution of the data is obtained. This recursive relabeling of phone categories results in an implicit modeling of the most significant contextual effects without resorting to heuristic decision criteria.

Instead of using a predetermined number and inventory of speech units, on-line methods employing selection of nonuniform units have also been tested. In this process, units for concatenation are determined at synthesis time by searching a large corpus to find the optimal sequence of phones embedded in natural corpus utterances that match the target utterance (see Fig. 12). Because the units are excised at synthesis time from optimal contexts, they are longer than prestored units and have correspondingly fewer join points. This has the dual advantage of ensuring appropriate coarticulation effects and at the same time reduces the number of potential discontinuities in the synthesized speech.

Because longer units decrease the number of concatenation points, it is commonly believed that they are superior to shorter units for concatenation. This is not necessarily the case. If phone-sized or subphonemic units are excised from ideal environments parallel in all relevant dimensions, then they can be concatenated without signal processing and with no noticeable discontinuities. The cost of finding such units is that the source corpus must be very large indeed.

Natural-Speech Synthesis

By extending the nonuniform principle to include prosodic contexts and phonemic contexts as selection criteria for synthesis units, the need for subsequent signal processing is almost eliminated, and the concatenated speech has the quality of the original high-fidelity recordings.

By labeling each phone in the corpus with its prosodic characteristics and with a simple phonemic label, an index of all phones in the corpus can be prepared as the basis of a selection process that minimizes discontinuities in both prosodic and spectral domains simultaneously.

In this method, two cost functions, a target cost and a join cost are simultaneously minimized by Viterbi search through a preselected number of candidate segments. Several instances of each unit are considered as potential candidates for concatenation, so the size of the database needs to be extremely large.

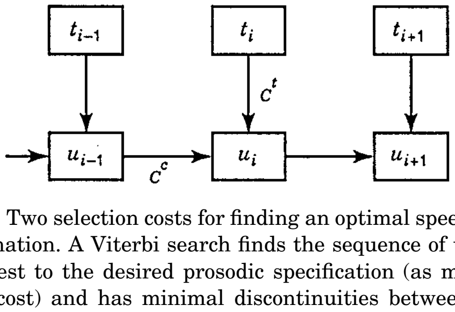


Figure 12. Two selection costs for finding an optimal speech segment for concatenation. A Viterbi search finds the sequence of units which that is closest to the desired prosodic specification (as measured by the target cost) and has minimal discontinuities between them (as measured by the concatenation cost).

Corpora for Speech Units

Speech corpora recorded for producing of diphone databases are very small. Each sentence typically consists of a fixed frame utterance in which a nonsense word is embedded, and two or three diphones are usually cut from each nonsense word. When the sentences are read, the speaker is instructed to keep the prosody as neutral as possible so that diphones concatenate easily and so that the prosody is uniformly modified.

Corpora for natural-speech synthesis, on the other hand, need to include as much prosodic variation as possible so that the component sounds cover all the required contexts. Currently available source corpora are limited at most to only a few hours of speech from any single speaker but have been produced for many voices and for several different languages.

Whereas the early corpora for concatenative synthesis were designed using greedy algorithms to include the smallest number of rich meaningful sentences that guaranteed full coverage of all segmental combinations, the resulting texts proved difficult to read fluently, and the ensuing tension in the voice resulted in less than optimal prosody for the source databases. More recently, it has been confirmed that continuous texts result in more natural prosody and also in a more natural voice quality because of fewer hyperarticulations in the reading.

TRENDS IN SPEECH SYNTHESIS

Trends in speech synthesis have followed developments in computer hardware and programming philosophy. In the eighties there were strong advances in statistical programming methods, and the development of neural networks, binary classification and recursion trees, and hidden Markov sequence modeling. These techniques were soon adopted for modeling the regularities in natural speech data instead of trying to emulate them using knowledge-based or heuristic methods. The development of these advanced learning algorithms also coincided with increases in computer storage capacity and in computing power and facilitated the corpus-based approaches.

The advantage of such mathematical modeling of speech is that it facilitates replication and verification of synthesis techniques that previously relied on careful hand tuning and were rarely easily generalizable. What is now developed for one speaker and one language can be ported with little effort to other speakers and other languages by simply substituting the appropriate speech corpora and rerunning the same learning algorithms.

Corpora as Knowledge Sources

Large speech corpora became available for synthesis research in the eighties and were originally used to analyze and train prosodic parameters, revealing tendencies in the data and providing test data against which trained models could confirm how an unseen utterance should be spoken in a given situation.

The same corpora were also used for source units of speech segments and prompted the development of concatenative speech synthesis, allowing close replication of the voice characteristics of the original corpus speaker and eliminating the necessity for modeling the intricate transitions between relatively steady states of individual speech segments.

Originally limited to diphones, the corpus-based synthesizers soon moved on to nonuniform source units, segments of speech waveform that range in size from subphonemic to multiword chunks, excised from longer samples of real speech and concatenated to produce novel utterances.

The development of corpus-based speech synthesis techniques in the nineties largely resulted from hardware and software developments in the computer world but considerably improved the quality of synthesized speech so that it is now capable (at times) of being mistaken for that of a human speaker.

Speech and Personality

Perhaps one reason for the slow take-up of speech synthesis technology is the fact that synthesizers can still portray only a small part of the information carried by a human voice. They lack personality, mood, emotion, and express only the minimum of focusing and emphasis, reducing the speech to an aural version of its text but losing much of the structural information.

Early corpus-based approaches to speech synthesis required signal processing to modify the prosody of the selected units. But more recently, because very large amounts of memory that accompanied multimedia computing developments are available, the use of extremely large corpora as a source of speech units has led to advances in segment selection techniques so that prosodically appropriate segments are selected and concatenated without recourse to potentially degrading modifications to the signal quality.

Emotion in Speech

Because signal processing is reduced in large-corpus concatenative speech synthesis, the style and characteristics of the input speech are preserved verbatim in the novel utterances, but to express different emotions or speaking styles, even larger corpora become necessary.

It is already common in speech recognition to use domain-specific models and grammars for reducing the perplexity of the recognition task by predicting the likely candidates from a context-limited range of candidates. Because speech synthesis methodologies have progressed in the past by adopting recognition developments, future concatenative synthesis can be expected to progress in much the same way, using domain-specific corpora to express appropriate speaking styles or emotions.

Whereas the challenge in creating voices for early synthesizers lay in modeling segments and their transitions and in predicting appropriate parameter tracks for modeling speech characteristics, current challenges are in collecting and annotating speech corpora of sufficient size and variety to allow modeling speaking styles and emotions appropriate for expressing finer distinctions of meaning. Rather than being reading machines for which not much need has been found, future synthesizers are more likely to take on the role of speaking machines and will be used to present on-line information interactively, thereby needing to express doubt and certainty in much the same way that humans do to offer a degree of confidence in the content of the utterance.

Multilinguality

Because of the increase in web-based information, we must also be prepared to process more multilingual input for speech synthesis. Currently synthesizers are being developed for all of the world's major languages, but few of these can process multilingual text or voice.

Recent developments in the automatic interpretation of spoken language offer potential application to multilingual speech synthesis. Indeed, they may even offer improved perception of synthesizer output by synthesizing voice in a foreign language. Figure 13 shows how the sequence of speech segments can be predicted language-specifically and then mapped in a language-independent way onto the voice of a speaker.

By mapping from the phone sequence and intonation predicted for one language onto the phone inventory and prosodic range of another, we synthesize a "foreign" language using the voice of a "native" speaker. This is sometimes necessary when processing a text containing words of more than one language to avoid the necessity of using two voices when there is actually only one source for the text. A side effect of this process is the perception that the native speaker is multilingual because few humans switch languages with such apparent fluency.

Mark-Up Languages

As a first step to giving speaking machines the range of expression they will eventually require, mark-up languages have been proposed (see Fig. 14). Using HTML-style annotation of the input text, they allow fine specification of focus, speed of speech, loudness of the voice, sex and age of the speaker, and so on, which the text analysis component usually cannot predict.

Using such annotated input text (see Fig. 14), the synthesizer adapts its generation to match the desired interpreta-

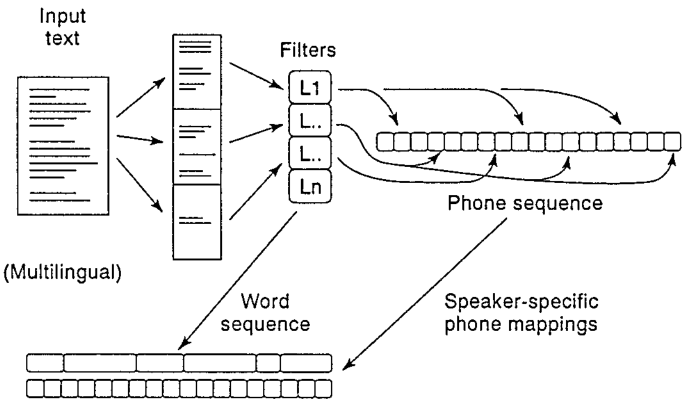


Figure 13. Processing multilingual text input. For each language that is represented in the input text (many of which can be identified by character coding alone), a language-specific filter converts the text into a pronounceable form to give a time-aligned multilingual phone sequence, which is mapped into the phones in the language of the synthesis speaker before further processing.

<table>
<tr>
<td>Range</td>
<td>Sets the "pitch range" of the intonation. A specification in one of the following formats:
• A positive floating-point number representing an absolute Hz value
• A percentage value higher or lower than the current. Thus, for N a floating point number, the following are legal specifications:
<table>
<tr><td>N%</td><td>N percent above current</td></tr>
<tr><td>+N%</td><td>N percent above current</td></tr>
<tr><td>–N%</td><td>N percent below current</td></tr>
</table>
• A descriptive term:
<table>
<tr><td>largest</td><td>largest available value for engine/speaker</td></tr>
<tr><td>large</td><td>reasonable large value for engine/speaker</td></tr>
<tr><td>medium</td><td>reasonable medium value for engine/speaker</td></tr>
<tr><td>small</td><td>reasonable small value for engine/speaker</td></tr>
<tr><td>smallest</td><td>reasonable smallest value for engine/speaker</td></tr>
<tr><td>default</td><td>reset to default value for engine/speaker</td></tr>
</table>
Default is 0%</td>
</tr>
</table>

Figure 14. Speech synthesis mark-up language allows specifying detail finer than the text analysis component can determine from the word sequence alone.

tion of the utterance. In conjunction with the Web Accessibility Interface included as part of HTML-4.0 (WAI) they enable an author to annotate the types of information in a text so that processors, such as search engines, can extract meaningful or relevant parts of a text or table for further processing.

A simple example of the need for such annotation can be seen in the display of a directory listing (%ls in UNIX or >dir in DOS, see Fig. 15), which usually defaults to an alphabetic ordering of file names in columns but is displayed on a computer screen as a table generated from left to right across the rows. When passing such generated text directly to a synthesizer, the visual ordering information is lost and the alphabetization appears random. Computer display is optimized for sighted people, not for speech.

EVALUATION OF SPEECH SYNTHESIS

To evaluate progress in speech synthesis methodologies, assessment techniques are required that provide measures related to human perception of speech. Currently web-based facilities generate randomized text sequences that have particular definable characteristics (see Fig. 16) but as yet no simple objective measure of the output of a speech synthesizer provides suitable quantification of its perceived quality.

Component Versus System

Part of the problem of evaluating synthesized speech is that so many components are involved, any one of which can be responsible for degrading perceived quality. If the text analysis is inadequate, then the prosodic prediction is not performed well. If the prosody is inadequate, then the selection of units is not appropriate. Within each of these major components are several subcomponents whose inadequate performance affects the end result, none of which is easy to evaluate in isolation.

Statistical methods learn the characteristics of the speech data well, but they are limited to objectively measurable features only. Because they only learn the data as presented, they have no concept of perceptual limens unless these are specifically represented in the training data. For example, predicting segmental durations is accurate to within a few milliseconds at the level of the phone, but compounding of prediction error at the level of the syllable can exceed the just noticeable difference and result in disrupting the perceived rhythm of an utterance. On the other hand, it has been shown that even a large error in predicting individual phone durations goes unnoticed if there is a corresponding error in the opposite direction in the predicted duration of a neighboring phone within the syllable. Therefore such objective measures are inadequate to quantify perceived quality in synthesis speech.

Naturalness Versus Intelligibility

Evaluation of computer speech has been focused in the past on measuring intelligibility rather than naturalness, based on the assumption that natural-sounding speech is of less importance. However, with the changing needs and ever-improving quality of synthesized speech, this assumption is open to challenge.

```
user@host% ls /
DB       data4    export       pcfs      usr
bin      dept1    home         prj       var
boot     dept2    homes        sbin      vmunix
data     dept3    kadb         server
data1    dept4    lib          sys
data2    dev      lost+found   tmp
data3    etc      mnt          tmp_mnt
user@host%
```

Figure 15. A directory listing shows file names arranged alphabetically in columns in vertical order, but the screen is actually written in rows from left to right, starting from the top. So if such a listing were to be sent directly to a voice output device, the ordering would appear meaningless.

1. The wrong shot led the farm.
2. The black top ran the spring.
3. The great car met the milk.
4. The old corn cost the blood.
5. The short arm sent the cow.
6. The low walk read the hat.
7. The rich paint said the land.
8. The big bank felt the bag.
9. The sick seat grew the chain.
10. The salt dog caused the show.

Figure 16. The first 10 sentences from the Haskins Anomalous test set designed to minimize semantic prediction when testing the comprehensibility of synthetic speech.

When segmental intelligibility was the biggest problem of parametrically generated computer speech, it was necessary to confirm that each sound had the required characteristics to perceived the words correctly. This prompted the creation of nonsense syllables and semantically anomalous sentences (see Fig. 17) for testing segmental intelligibility, but because the speech was generated from the concatenation of segments taken from human speech, the problem of intelligibility became one of naturalness.

It has been argued that there is no need for computer speech to emulate that of humans, and even that there may be a need to distinguish computer-generated speech from that produced by people for reasons of reliability. If this is the case, then naturalness is not a relevant criterion. But when listening to speech under degraded listening conditions, such as in a moving car or in a noisy room, the rendundant information in a real speech signal allows for improved intelligibility.

Whereas intelligibility is easy to measure, using dictation tests, minimal-pair segmental confusion tests, and comprehension tests, etc., the perception of naturalness is not so readily quantifiable. Because of increasing reality in synthetic speech, evaluation can become a matter of personal preference. Television presenters, such as news anchors, attract audiences because of their personalities, much of which are portrayed in the voice. Voice quality and speaking style are volatile attributes, as listening to the news broadcasts of even a few years ago illustrate. If we were to judge synthesizers as we judge newscasters, then the reliability of their results would be quickly called into question. Perhaps this is a matter best left to market forces.

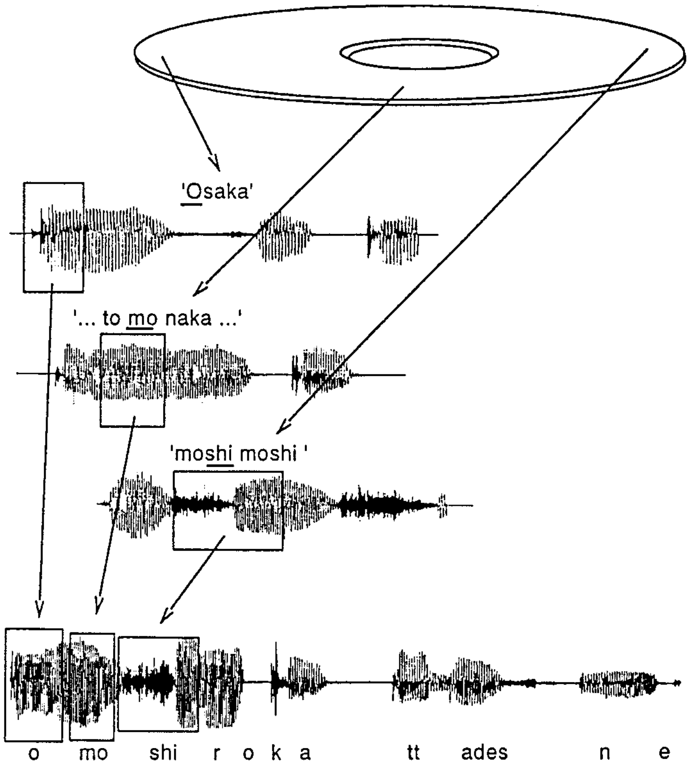


Figure 17. Selecting waveform segments from a large database of speech and concatenating them to create novel utterances. In this example, the word omoshirokattadesue (Japanese for "wasn't that interesting") is formed by joining the /o/ from Osaka, the /m/ and the / o/ from "nantomonakatta", the /sh/ and the /i/ from "moshi-moshi, and so on. The computer searches the database for single phonemes but retrieves longer sequences if they are naturally contiguous *and* match the desired prosodic target.

RESPONSIBILITY

Because of the increasing naturalness of concatenative speech synthesis systems, it is now possible to replicate the voice and speaking style of a person so that others believe that the synthesized utterance was actually spoken by the original speaker of the source database.

Techniques exist for "watermarking" electronic signals so that their source can be ascertained even after repeated copying or duplication. If use is to be made of convincingly naturally sounding synthesis, then it might be in the best interests of both the corpus speaker and the synthesis developer to ensure that there is a method of tracing the source of a synthesized utterance to prevent misuse of the technology.

Similarly, in many countries, the voice of a person is not yet subject to copyright. The sounds of a voice are considered of equivalent status with individual words in a text or colors in an image, and it is generally only the original and novel sequences or combinations of such basic units that are protected under law. When voice reproduction was limited to tape recordings, this view still provided a degree of protection to the speaker, but the concatenation of single phone-sized sounds realistically may be legal even without the permission of the speaker. This is not a desirable situation.

Finally, the point of minority representation is raised here. All of the major industrial nations have produced speech synthesizers for their own languages, but (with some very notable exceptions) there is little general support for synthesis of minority languages. Because language is closely connected with cultural identity, developers of the technology should take responsibility to ensure that there is fair coverage of as many languages as possible by encouraging the collection of corpora to further the study of such languages and by making their systems less language-dependent and more generic.

FURTHER INFORMATION

For those interested in obtaining further information about computer speech synthesis, this section offers a list of sources on paper and via the internet.

Books

Several books are devoted to speech synthesis. Perhaps the first and certainly required reading from a historical point of view, is the seminal MITalk book: John Allen, Sharon Hunnicut, and Dennis H. Klatt, *From Text to Speech: The MITalk System,* Cambridge University Press, 1987.

The quadrennial ESCA tutorial workshops on speech synthesis also produce books of selected papers, complied in greater detail after the workshop, that are considered required reading for researchers in the field. They document the major developments and changing focus of the technology. The first in this series was *Talking Machines, Theories, Models and Designs,* Gerard Bailly and Christian Benoit eds., Elsevier: North Holland, 1992, and the second *Progress in Speech Synthesis,* edited by J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, (Springer, 1996). The third

Speaking Machines, (Springer, 1999) will be published simultaneously with the present volume.

Other recent books on speech synthesis include

- T. Dutoit, *An Introduction to Text-to-Speech Synthesis,* Dordrecht: Kluwer Academic Publishers, 1997, ISBN 0-7923-4498-7.
- D. O'Shaughnessy, *Speech Communication: Human and Machine,* Addison-Wesley series in Electrical Engineering: Digital Signal Processing, 1987.
- V. van Heuven and L. Pols, (eds.), *Analysis and Synthesis of Speech,* Mouton de Gruyter, 1993.
- I. H. Witten. *Principles of Computer Speech,* London: Academic Press, 1982.
- W. B. Kleijn and K. K. Paliwal (Eds.), *Speech Coding and Synthesis,* Amsterdam: Elsevier, 1995.

Journals

Articles relevant to speech synthesis technology appear in the *Journal of the Acoustical Society of America, Computer Speech and Language, Speech Communication, Phonetica,* the *Journal of Phonetics,* and the *Journal of Speech Technology.*

References to a large number of speech synthesis papers are stored under the COCOSDA archives at http://www.itl.atr.co.jp/cocosda/synthesis. These are updated periodically by scientists working in all fields related to speech synthesis, and although not an authoritative source, provide a good background to the sorts of titles, conferences, and journals that attract synthesis researchers.

Conferences/Workshops/Societies

Perhaps the biggest international conference of the speech and signal processing community is the annual *International Conference on Speech and Signal Processing* (ICASSP), but the time devoted to speech synthesis at these meetings is very brief and is restricted mainly to the signal processing aspects of the technology.

Two biannual international conferences that are widely attended by researchers interested in speech synthesis are the *International Conference on Spoken Language Processing* (ICSLP) and *Eurospeech* (which is European in name only). These meetings are attended by linguists, psychologists, engineers, and educators.

Two other large international conferences, the *Association of Computational Linguists* (ACL) and the *Meeting of Computational Linguists* (Coling) are recently incorporating more speech technology presentations in their proceedings, reflecting the growing trend for integration of speech and language technologies.

The *European* (again, in name only) Speech Communication Association (ESCA), offers speech synthesis tutorial workshops every four years and has a special interest group (SIG) devoted to disseminating information related to speech synthesis.

Synthesis on the Internet

The International Coordinating Committee on Speech I/O Databases and Assessment (COCOSDA), offers a synthesis web page that attempts to correlate worldwide information as part of its information gathering preparatory work for Standards formation. See www.itl.atr.co.jp/cocosda for more details.

There are many World Wide Web sites that archive speech synthesis samples and offer interactive access to synthesis systems. Visitors can submit texts and listen to the synthesized results interactively. Following is a selection:

- HADIFIX German Speech Synthesis: http://asl1.ikp.uni-bonn.de/tpo/Hadiq.en.html
- Institute of Phonetic Sciences: http://fonsg3.let.uva.nl/IFA-Features.html
- Museum of Speech Analysis and Synthesis: http://mambo.ucsc.edu/psl/smus/smus.html
- Web sites concerning Speech: http://ncvs.shc.uiowa.edu/misc/other-sites.html
- ICG Grenoble's "exemples sonores": http://ophale.icp.grenet.fr/ex.html
- Speech Synthesis at ICP Grenoble: http://ophale.icp.grenet.fr/home.html
- Microsoft's Speech Synthesis Project: http://research.microsoft.com/stg/ssproject.html
- The MBROLA project homepage: http://tcts.fpms.ac.be/synthesis/mbrola.html
- MBROLA: Free Speech Synthesis Project: http://tcts.fpms.ac.be/synthesis/modelcmp.html
- Lector (Spanish): http://www.angelfire.com/biz/lector
- AT&T Advanced Speech Products Group: http://www.att.com/aspg/
- Lucent Technologies Bell Labs Text-to-Speech: http://www.bell-labs.com/project/tts/
- ORATOR from Bellcore: http://www.bellcore.com/ORATOR/
- BeSTspeech from Berkeley Speech Technologies: http://www.bestspeech.com/weblang.html
- Centigram's TruVoice: http://www.centigram.com/centigram/TruVoice/index.html
- The Birmingham Speech Synthesis Museum: http://www.cs.bham.ac.uk/jpi/synth/museum.html
- Speech Synthesis from OGI: http://www.cse.ogi.edu/CSLU/research/TTS
- CSLU Speech Synthesis Research Group: http://www.cse.ogi.edu/CSLU/research/TTS/
- Lyricos: http://www.cse.ogi.edu/CSLU/research/TTS/research/sing.html
- Festival Speech Synthesiser: http://www.cstr.ed.ac.uk/projects/festival.html
- EUROVOCS: http://www.elis.rug.ac.be/ELISgroups/speech/research/eurovocs.html
- Eloquent Technology, Inc. A Speaking Web Site: http://www.eloq.com/
- TTS from Duisburg: http://www.fb9-ti.uni-duisburg.de/demos/speech.html
- First Byte Text-To-Speech HOME PAGE: http://www.firstbyte.davd.com/
- Haskins Laboratory WWW Site: http://www.haskins.yale.edu/Haskins/MISC/special.html
- Musee sonore de la synthese de la Parole en francais: http://www.icp.grenet.fr/exemples-synthese/ex.html
- HADIFIX: http://www.ikp.uni-bonn.de/tpo/Hadifix.-en.html

- Stuttgart's Synthesis Collection: http://www.ims.uni-stuttgart.de/phonetik/gregor/synthspeech/examples.html
- CHATR (ATR's multilingual speech synthesis system): http://www.itl.atr.co.jp/chatr
- BT Laboratories—Text-to-Speech: http://www.labs.bt.com/innovate/speech/laureate/
- NTT's Japanese synthesis: http://www.ntt.co.jp/japan/japanese/
- Infovox: http://www.promotor.telia.se/infovox/index.htm
- AT&T Research Voices: http://www.research.att.com/cgi-bin/cgiwrap/mjm/voices.cgi
- Microsoft Speech Research: http://www.research.microsoft.com/research/srg
- Pavarobotti: http://www.shc.uiowa.edu/fun/pavarobotti/pavarobotti.html
- IBM Voicetype: http://www.software.ibm.com/is/voice-type
- Apple's PlainTalk: http://www.speech.apple.com/speech/ptk/ptk.html
- Kungliga Tekniska Hogskolan: http://www.speech.kth.se/info/software.html
- Multimodal Speech Synthesis from KTH: http://www.speech.kth.se/multimodal/
- Speech Toys: http://www.speechtoys.com/spchtoys/spsyn.html
- SoftVoice, Inc.: http://www.text2speech.com/
- SVOX from TIK, ETH in Zurich: http://www.tik.ee.ethz.ch/cgi-bin/w3svox
- WebSpeak: http://www.tue.nl/ipo/hearing/web-speak.htm
- Bibliography Phonetics and Speech Technology: http://www.uni-frankfurt.de/ifb/bib-ngl.html
- Say: http://wwwtios.cs.utwente.nl/say/
- Eurovocs Multilingual Speech Synthesis: http://www.elis.rug.ac.be/ELISgroups/speech/research/eurovocs.html

Mailing Lists

A FAQ file of Frequently Asked Questions about speech synthesis is archived under comp.speech, a commonly subscribed list that has mirror sites at www.itl.atr.co.jp, squid.eng.cam.ac.uk, www.speech.cs.cmu.edu, and www.speeech.su.oz.edu.

There are two mailing lists devoted to speech synthesis: synth@bham.ac.edu and cocosda@itl.atr.co.jp. Both are administered under the automatic majordomo mailing list software and can be joined by sending email to the user majordomo at the same address as the mailing list, with the words "subscribe list-name your-name." If the messages prove too frequent or of insufficient interest, you can unsubscribe by sending email to the list address with the words "unsubscribe list-name your-name". Archives are usually kept of previous messages and they are a useful source of information to researchers wanting to learn more about the technology.

NICK CAMPBELL
ATR-ITL

SPEECH TECHNOLOGY APPLICATIONS. See SPEECH PROCESSING.

SPEED MEASUREMENT. See VELOCIMETERS.

SPICE. See CIRCUIT ANALYSIS COMPUTING.