One way to gain an understanding of natural spoken language is to derive a constructive theory of it by building a language engine. The more nearly this artificial language engine simulates human linguistic behavior, the more of language it may be said to explain. A constructive theory of language must have two important features. It must capture known linguistic structure and it must express this structure in an elegant and computationally tractable mathematical framework.

Complete constructive theories of language do not yet exist. However, interesting ones do exist and there is reason to hope that they will improve. This paper describes the state-of-the-art in automatic speech synthesis and speech recognition and explains some of the mathematical models on which their underlying theories rest.

1. Introduction

The motivation for this article is the need for engineers and linguists to collaborate. In particular, the topic on which the two disciplines have mutual interests is spoken dialog between humans and machines. It is immediately apparent that the construction of machines that produce and understand natural spoken language is a holy grail for electrical and computer engineers. It is further obvious that the science of linguistics has a great deal to say on the subject. Linguistics offers not only technical advice on precisely what human-machine dialog entails, but also theoretical considerations of the architecture of the human language engine. This view of the collaboration might be interpreted as simply a consultation in which engineers have much to learn and little to teach. Fortunately, the collaboration may legitimately be seen as mutually beneficial. The very thought of building a machine capable of speaking and understanding speech is nothing less than a constructive theory of language. The more nearly the attempt succeeds, the more light it sheds on the linguists’ central questions.

Indeed, there have been some successful collaborations between linguists and engineers on this very topic. Unfortunately, these collaborations have been fragile and thus have not achieved their promise. Here we explore some of the joint efforts and suggest ways in which they can be less fragile and more effective.

With respect to the problem of human-machine communication by voice, engineering and linguistics make the following contributions. Engineering offers
the mathematics and physics required to make quantitative models of the processes involved in speech communication.

Linguistics provides detailed qualitative descriptions of the structure and usage of language.

While these insights are certainly necessary to a scientific understanding of language, they cannot be applied blindly. Engineers must realize that mathematical models, no matter how elegant and sophisticated they may be, are useful only to the extent that they capture the essential structure of the phenomenon under consideration, in this case, spoken language. On the other hand, the linguist's taxonomy of structures and usages, rules, and examples, no matter how exhaustive, are merely anecdotal evidence and, as such, of limited value unless they can be embedded in a rational computational framework.

These characterizations will, no doubt, be criticized as simplistic and stereotypical.

Linguistics is sometimes rigorous and quantitative. Nor is Engineering always brute force calculation. The best way to see some of the subtleties is to examine some case histories. In particular, it is useful to consider the state-of-the-art in text-to-speech synthesis (TTS) and automatic speech recognition (ASR). As used here, TTS refers to the process of generating an acoustic speech signal without regard for its meaning. The generated speech should be intelligible to a human listener, sound natural, and convey useful information, all despite the fact that the generation process has no means to represent semantics.

Similarly, ASR is intended to refer to the inverse process, that of transcribing speech into text without regard for meaning. It is expected that the accuracy of transcription should be nearly perfect, independent of speaker and topic.

It is not at all certain that these problems, as stated, can be solved. In fact, it is not even clear that, were they solved, the solutions would be of any practical value. The debate, of course, hinges on the absence of semantic processes in both cases. Some research efforts have acknowledged these difficulties and have addressed the more complex problems of speech synthesis from concept and automatic speech understanding. For the purposes of exploring the interaction of engineering and linguistics, it is not necessary to consider these additional complications.

2. Speech Synthesis

The best example of a collaboration of linguistics and engineering is that of speech synthesis from text. The state-of-the-art in TTS is quite advanced. Speech synthesizers can read absolutely any text with a high degree of intelligibility in several different voices. The naturalness of the voices is quite good but would never be mistaken for a human voice by even the most naive of listeners. Strangely enough, TTS has been less of a commercial success than its companion technology (ASR), even though the latter is technically far less proficient.
Speech synthesis is far more intuitively comprehensible than is ASR. The generation of sound was well understood by ancient musicians and the analogy of musical instruments to the vocal apparatus led, as early as the 18th century, to mechanical speaking devices (von Kempelen 1791).

As for translating the written word to a sequence of sounds, anyone taught to read phonetically finds the concept quite natural.

The mechanical embodiment of these ideas is shown in the diagram of Figure 1 (van Santen & Sproat 1998). It is understood that all of the processes indicated in the figure are carried out on a digital computer.

![Text-to-Speech Synthesis Diagram](image)

**Figure 1.**

The first five of the processes account for the transliteration of standard orthography into its phonetic equivalent. These processes taken as a whole represent, very possibly, the best compendium of the linguist's knowledge of phonetics, phonology, phonotactics, morphology, and prosody. The last box represents the engineer's best understanding of the physics of sound generation in the human voice apparatus.

The details of the operation of this system are instructive. The conversion of graphemes to phonemes, although it spans several levels of linguistic structure, is virtually monolithic, namely table look-up. The tables are large pronouncing dictionaries. Thus text normalization is simply a list of abbreviations, acronyms,
counting numbers and non-alphameric symbols along with their usages and pronunciations.

In text normalization as well as lexical access and syntactico-semantic analysis, there are always ambiguities that affect pronunciation. For example, Dr. can be pronounced as doctor or drive (as in an address). The word bass will be pronounced differently when it means a fish or a stringed instrument. And, of course, read will be pronounced differently when it is present or past tense.

All of these issues are resolved by the same mechanism, concordances based on the information-theoretic property of mutual information. The mutual information between two words is the negative binary logarithm of the ratio of their joint probability to the product of their prior probabilities. Thus when two words are likely to appear together they have high mutual information. The words with which a given word has high mutual information determine its usage, hence its pronunciation and/or its prosodic features. For example, if Dr. appears with a numeral it should be pronounced drive. If bank appears with river, it should be un-stressed. There are, of course, vast numbers of such ambiguities in natural language. The mutual information coefficients needed to resolve them are computed exhaustively from large textual corpora.

In addition to the primary lexical and syntactico-semantic analysis described above, there is a secondary syntactic analysis required. This is a crude parse used to find phrase boundaries which, in turn, are used to assign pitch contours and accents. Note that a full syntactic parse into parts of speech is not required.

The phonetic and phonological analyses are also largely accomplished by table-look-up. First, however, a morphological analysis must be performed to make the table-look-up more efficient. The rule-based morphological analysis decomposes words into their base forms and their inflections thereby reducing the number of entries needed in the pronouncing dictionary.

Unfortunately, it is not practical to store the pronunciations of all morphemes. To account for this, two alternate methods of phonetic analysis are provided. The first is to use a pronunciation of a morpheme that rhymes with the missing one. The second is a set of letter-to-sound rules. Such rules are not reliable and thus are used only as a last resort.

Once the phonetic pronunciation has been determined from the dictionary, phonological analysis is performed. In order to understand how this is accomplished as a table-look-up, it is first necessary to recall that the acoustic/phonetic units are actually sequences of allophones called polyads. There are about 2500 such units stored as sequences of frames of linear prediction coefficients excised from natural speech (Olive et al. 1998). The phonology is implicit in the selection of the units. That is, the units are selected to give the broadest coverage of the phonology of the entire language. When synthesizing fluent speech, a morpheme is realized by selecting the sequence of polyads that most closely matches its phonological context. The selection of the inventory of polyads is carried out automatically by an optimal algorithm.
Finally, after a sentence or paragraph has been analyzed with respect to phonetics and phonology, the suprasegmental prosodic features are superimposed. That is, each of the frames, i.e. LPC vectors, of each polyad is marked immediately with pitch, intensity, and duration. Acoustic synthesis follows by conventional LPC methods. The parameters of the synthesizer may be adjusted to produce different stereotypical voices.

It is appropriate to comment here about the mathematics of the acoustic synthesis procedure. The method of linear prediction was originally derived for the purpose of analyzing time series such as sunspot activity (Yule 1927). When so used, linear prediction is nothing more than brute-force curve-fitting with no underlying model. However, it can be shown (Wakita 1973) that the abstract mathematics has a very interesting interpretation, namely, it is the solution to the linear wave equation in a hard-walled tube of varying cross-sectional area. Here, then, is an excellent example of mathematical analysis working well in linguistics, because it captures a fundamental property of the phenomenon under consideration.

The method of synthesis described above is, for obvious reasons, called concatenative synthesis. One might be tempted to object that it is not true synthesis, because it is really just a sophisticated recording device which reproduces speech as sequences of brief stored segments. An alternative method called articulatory synthesis addresses this criticism by synthesizing speech directly from the physics of an articulatory model (Figure 2) (Coker 1976) without any prerecording of any kind. Using the very same linguistic analysis as outlined above, as you might guess, the resulting synthetic speech, while intelligible, is of far worse quality than that generated by concatenative methods.

Figure 2.
In summary, then, speech synthesis can be accomplished by a careful, detailed, exhaustive encapsulation of linguistic knowledge in ‘dictionaries’ of various kinds constructed by well-chosen mathematical analysis. This significant technical accomplishment is the result of an ideal collaboration between engineers and linguists. Based on that joint accomplishment, one may dare to hope that the problem of articulatory synthesis will also be solved.

3. Speech Recognition

The practice of speech recognition does not present so cheerful a picture as does speech synthesis. The state-of-the-art is not nearly as advanced and the interaction of engineering and linguistics not nearly as cooperative. The result is best described as an engineering tour-de-force with a condescending tip of the hat to linguistics. Still, the status quo is instructive.

First, we must admit that for machines, as for people, listening is harder than talking (both literally and figuratively). In the case of synthesis, we need only produce one voice, whereas in recognition we must accept any voice. In the earliest work on recognition of acoustic patterns, little attention was paid to the high degree of variability in the speech signal. In fact, quite the opposite was true. The foundation of ASR, which lies in the seminal work of Visible Speech (Potter, Kopp, & Green 1968), is essentially a catalog of the ‘invariant’ spectrographic features of speech. The early electronic devices for ASR were based on capturing these ‘reliable’ features (Dudley & Balashek 1958). However, in the 1960’s, the emphasis shifted from cataloging and recognizing invariant features to characterizing speech as a stochastic process and using highly developed mathematical techniques for detection, estimation, and classification to analyze it (Sebestyen 1962). This transformation set up an almost insurmountable barrier between linguists and engineers that stands to this very day. Little information flowed across this barrier in either direction. However, the descriptive aspects of linguistics were accessible to some engineers, while the rigorous mathematics of engineering were of little concern to linguists. Happily this situation is now beginning to change.

![Diagram of speech recognition process]

Figure 3.
Skipping over the early history of ASR, let us look at the modern state-of-the-art. Today, large vocabulary recognition of fluent speech is accomplished by systems of the architecture shown in Figure 3. The interesting thing about this diagram is that linguists could have drawn it two or three decades ago. Unfortunately, they had no tools with which to implement it. The earliest attempts at an implementation were based on compiling an exhaustive list of rules for acoustic/phonetics, phonology, phonotactics, morphology, and syntax. These rules were applied by an ad hoc logical mechanism and followed by another ad hoc decision strategy to choose the best transcription for the utterance. The basic strategy is outlined in Newell et al. 1973, but no working version of the proposed system was ever constructed. In the absence of a rational mathematical framework, no amount of linguistic knowledge, regardless how detailed and comprehensive, can enable transcription of fluent speech. The problem is one of combinatorics. A large collection of heterogeneous rules is required. The rules have significant interactions with each other. The number of dependencies amongst the rules grows exponentially with the size of the rule set. No ad hoc procedure can ever be designed to apply and test these rules in an optimal, yet computationally efficient, way. And so, the early programs failed with linguists often blamed.

In the early 1970’s, the mathematical technique known as Hidden Markov Modeling was applied to speech recognition (Baker 1975, Jelinek 1976). The mathematics was known a decade earlier but, once again, it is especially appropriate to speech analysis because it naturally captures many aspects of linguistic structure.

Unfortunately, the engineers and mathematicians who applied the methodology to ASR, did so in a very clumsy way which uses the HMM to capture only the statistical structure of the speech signal. The early implementations of the HMM rested on the observation that speech is a quasi-stationary process, i.e., one in which the statistics of the signal are nearly constant over intervals of from tens to hundreds of milliseconds in duration. The hidden states of the HMM were therefore identified with the quasi-stationary regions. In order to force all aspects of linguistic structure to conform to this single notion, the system architecture of Figure 3 was revised as shown in Figure 4, in which all levels of linguistic structure are combined uniformly into a single vast HMM.

In order to accomplish this compilation, one assumes that all phonetic units (phonemes) have three parts, an onset, a steady state or target, and a decay. These are represented by a three-state non-ergodic HMM. It is further assumed that phonology is accounted for by triphonic variation, that is each phonetic unit is influenced only by its immediate predecessor and successor. A different HMM for each phonetic unit is generated for each such phonetic environment. Finally, phonotactic structure is imposed by allowing only those sequences of phonetic units that appear in valid word sequences. A valid word sequence is one whose trigram probability is non-zero.
Figure 4.

Overall block diagram of subword unit based continuous speech recognizer.

The result of these assumptions is a huge HMM with millions of parameters whose values are automatically estimated from hours of unlabeled (i.e., unsegmented) speech of many different speakers. It is another engineering tour-de-force that such a model can be built.

Even more impressive is the fact that the method works vastly better than its early rule-based ancestors. In fact, for vocabularies of tens of thousands of words, fluently read speech of almost any speaker (i.e., native speaker of American English) will be transcribed with 90% accuracy. Considering that the transcription is performed without any knowledge of the meaning of the utterance, this result is remarkable.

4. Conclusion

As noted earlier, this method works because it is based on linguistic structure, albeit highly oversimplified. The lesson that engineers learned from their success was that rudimentary linguistics embedded in a powerful mathematical framework is all that is required. Linguistic subtleties can be safely ignored.

A more interesting implementation of Figure 3 captures a great deal more linguistic reality. Based on the Cave-Neuwirth experiments (Cave & Neuwirth 1980), it uses the more complex HMM shown in Figure 5 (Levinson 1986). This model is ergodic, with each state corresponding to a unique phonetic unit (allophone). Phonotactics is much more faithfully represented by the state transition matrix and segmental duration is explicitly represented by appropriate probability density functions.

Furthermore, the system retains the modularity implicit in the diagram by using separate but mathematically optimal algorithms for lexical access and pars-
ing, the latter based on a formal grammar of English. One encouraging result of this method is that, unlike the system of Figure 4, this system can produce phonetic transcriptions of words not in the lexicon.

Yet, for all its linguistic sophistication, this method yields the same performance as the single HMM technique. True enough, this system is more amenable to the addition of linguistic structure, but its observed behavior is not appreciably better, even though from a psychological and linguistic perspective, it is much more natural.

![Diagram](image)

Figure 5.

This disappointing fact is easily explained. Neither system displays anything even remotely approaching human linguistic abilities. There is no morphological
analysis, no prosodic analysis, and syntax is taken to mean only word order. Thus, there is no bridge to semantics, let alone an actual semantic analysis.

Thus there is hope for the future. Good linguistic theories for all the missing structures appropriately represented in a computationally rigorous, but tractable model will lead to the holy grail of automatic speech recognition at human-like levels of performance.

I have tried, over the past many years, to effect this kind of a research effort with little success. The impediments seem to be the following. Engineers are very proud of their recent accomplishments. Most feel that incremental improvements to existing systems will ultimately produce the desired result. I have argued against this sentiment (Levinson 1994) but it is hard to do in the face of the failure on the part of many engineers to recognize how amazing, robust, complex, and versatile natural language truly is.

On the other hand, linguistic theories often seem rather esoteric relative to the practical questions engineers ask. Furthermore, it often appears that linguistic theories stand or fall on the basis of carefully contrived anecdotal examples. As insightful as these may be, an ASR system requires an exhaustive collection of such theories to completely cover all linguistic phenomena. This often requires long and boring labor. Even after all the work is done, some parsimonious representation (probabilistic) must be devised.

And then, supposing progress could be made toward building an ASR machine. Many linguists would argue that such a machine would be a very narrow expression of linguistic theory and would not address the most important questions linguistics poses. I, of course, am a strong advocate of ‘constructive’ linguistic theories, and it is my fervent hope that some deep and honest introspection can reconcile these differences to the advantage of both disciplines.

REFERENCES


Levinson: Human-machine communication by voice


Session III:
Curriculum Design for Linguistic Purposes

Chair: Ladislav Zgusta

Saturday, 31 October 1999
1:00 p.m. - 5:00 p.m.