Speech recognition through spectrogram matching

Frances Ingemann and Paul Mermelstein

Citation: The Journal of the Acoustical Society of America **56**, S27 (1974); doi: 10.1121/1.1914090
View online: https://doi.org/10.1121/1.1914090
View Table of Contents: https://asa.scitation.org/toc/jas/56/S1
Published by the Acoustical Society of America

based on dynamic programming. The algorithm finds a sequence W_0 such that $P(W_0|A,L)$ is the maximum achievable over all possible sequences in the language L. The system has been tested on several languages, each representing a specialized interactive computer task, with vocabulary sizes ranging from 24 words to 195 words. In a series of experiments, each consisting of about 20 sentences from a given language, the system correctly recognized between 85% and 95% of the words.

2:30

N3. Speech recognition through spectrogram matching. Frances Ingemann (Haskins Laboratories, New Haven, Connecticut 06510, and Linguistics Department, University of Kansas, Lawrence, Kansas 66045) and Paul Mermelstein (Haskins Laboratories, New Haven, Connecticut 06410)

In order to assess human analysis of acoustic data before attempting such analysis by machine, a series of experiments was conducted in which subjects were asked to match spectrograms of continuous speech to reference spectrograms of the same words. Although error rates varied with sentence difficulty and size of vocabulary, comparison of the matches shows greater agreement in phoneme segments than other experiments have obtained in phonetic transcriptions of unknown utterances without semantic or syntactic processing. Accuracy in matching can be further improved by feedback in the form of spectrographic representation of a sequence of tentative matches spoken as if they made up the unknown utterance. Automatic matching of word- or syllable-sized acoustic patterns may provide a more accurate phonemic input to the syntatic—semantic component of a speech recognition system than other methods so far attempted. [Research supported by the Advanced Research Projects Agency, Department of Defense.]

2:45

N4. An overview of the Lincoln Laboratory speech recognition system. J.W. Forgie, D.E. Hall, and R.A. Wiesen (MIT Lincoln Laboratory, Lexington, Massachusetts 02173)

The Lincoln speech recognition system is capable of recognizing spoken sentences made up of words drawn from a limited vocabulary and constrained to conform to a context-free grammar. Speech input is taken from a close-talking, noise-cancelling microphone in a relatively noisy computer room. The speech is digitized and subjected to a detailed acoustic—phonetic analysis [C.J. Weinstein *et al.*, Proc. IEEE Symposium on Speech Recognition (April 1974), pp. 89—100] which produces a string of acoustic phonetic elements (APELs). The APEL string is scanned by a linguistic module which attempts to find and score candidate sentences which satisfy the syntactic and semantic constraints of the grammar and which are composed of words having acceptable matches between phonemic dictionary spellings and APEL representations. Typical sentences of 3- to 4-sec duration require the order of one minute of computer processing. Tests have involved vocabularies of 125 to 500 words, grammar of varying complexity, many speakers (both male and female), and several hundred test sentences. [This work was sponsored by the Advanced Research Projects Agency of the Department of Defense.]

3:00

N5. Parsing and word matching in the Lincoln Laboratory speech recognition system. D.E. Hall and J.W. Forgie (MIT Lincoln Laboratory, Lexington, Massachusetts 02173)

The Lincoln system is designed to recognize sentences conforming to a context-free grammar. Parsing is guided by a heuristic evaluation function which combines individual word scores into parse path scores. The word scores measure the degree of correspondence between phonemic dictionary spellings and the results of an acoustic—phonetic analysis of the input sentence. Word scoring is based on two computer-generated scoring matrices derived from confusion statistics gathered from 113 sentences. The synchronization problem resulting from missing and spurious segments is simplified by first aligning the vowels (sometimes in more than one combination). The effect of phonological rules is handled by inserting optional phonemes in the dictionary and flagging others as possibly missing. Other rules, which are specific to Lincoln's front-end analysis, predict spurious segments and the effect of neighboring semivowels on vowel classification. [This work was sponsored by the Advanced Research Projects Agency of the Department of Defense.]

3:15

N6. An evaluation of the Lincoln Laboratory speech recognition system. R.A. Wiesen and J.W. Forgie (MIT Lincoln Laboratory, Lexington, Massachusetts 02173)

The current version of the Lincoln speech understanding system employs a vocabulary of about 250 words with a grammar permitting some five million sentences. The acoustic—phonetic and linguistic analyses were developed using a large corpus of data, including many sentences from this grammar. The primary purpose of the research reported here is to permit an evaluation of the system, independent of any input data used in developing the system. Six male subjects each formulated and spoke 25 sentences using charts depicting legal constructs. The processing time for each sentence was about one minute, after which the system displayed what it thought was spoken. Somewhat later, each subject returned to repeat the list of the sentences which he had formulated earlier. Without any adjustments in the system per speaker, about 50% of the sentences were completely correctly recognized for each subject in both situations. This statistic gives very little insight into system's performance, and data analyses will be presented that attempt to portray the strengths and weaknesses of the system. Major emphasis will be given to modifications in acoustic—phonetic analysis suggested by the data. [This work was sponsored by the Advanced Research Projects Agency of the Department of Defense.]

3:30

N7. Comparison of two speech understanding systems. B.T. Lowerre (Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213)

This paper presents the results of a comparison of two speech systems developed at Carnegie-Mellon University: Hear—Say—1 (1) and Dragon (2). Hear—Say—1, which consists of cooperating but independent knowledge sources, utilizes a best first search technique of the syntatic grammer. The search terminates when the evaluation of a completed proposed sentence achieves a heuristic threshold. In this way only the most likely paths are searched, thus achieving a fast though not always accurate recognition. Dragon utilizes a combined network of grammar and phonetic spellings for a generative Markov process. This allows Dragon to search all possible paths in parallel in an amount of time that is linear with utterance length. Dragon obtains a higher accuracy of recognition but with a higher computational overhead. Both systems have been tested on the same five sets of data consisting of 102 utterances and 564 words. Hear—Say—1 correctly identifies about 60% of the words in about 5 to 30 times real time, depending on the number of wrong paths searched. Dragon correctly identifies about 85% of the words in an almost consistent 50 times real time. The types of errors produced by both systems are discussed. [(1) Written by D.R. Reddey, L.E. Erman, R.D. Fennell, and B.T. Lowerre, (2) Written by J.K. Baker.]