AUGUST 11 2005

Perplexity—a measure of the difficulty of speech recognition tasks FREE

F. Jelinek; R. L. Mercer; L. R. Bahl; J. K. Baker

J. Acoust. Soc. Am. 62, S63 (1977)
https://doi.org/10.1121/1.2016299

the "buzziness" and lack of naturalness perceived in the resulting synthesized speech. We propose a new source model, which combines both pulse and noise sources in a novel way. Based on the observation that spectra of voiced speech sounds (e.g., voiced fricatives and even certain vowels) exhibit devoiced or incoherent high-frequency bands, the model divides the spectrum into a low-frequency region and a high-frequency region, with the pulse source exciting the low region and the noise source exciting the high region. The cutoff frequency that separates the two regions is adaptively varied in accordance with the changing speech signal. We present the advantages of the proposed model over the pulse/noise model, and describe a method for implementing it. Synthesis experiments conducted using the above model with manually extracted cutoff frequency data indicate the power of the model in almost entirely eliminating the "buzzy" quality. [Work sponsored by ARPA–IPTO.]

2:30

BB6. Testing recognition of computer-generated speech with elementary school children. Robert Laddaga and W. Sanders (Institute for Mathematical Studies in the Social Sciences, Ventura Hall, Stanford University, Stanford, CA 94305)

Recently, two experiments were performed on recognition of computer-generated speech. One experiment tested recognition of letter sounds by first graders, the other tested fifth graders' recognition of consonant sounds in monosyllabic words. The systems tested were the Allen/Klatt system [J. Allen, 1977 IEEE Int. Conf. ASSP, p. 579 (1977)], the Votrax VS6 and ML1 systems (all phoneme synthesizers), and the Institute's MISS system (linear predictive compression of recorded speech) [Sanders and Laddaga, J. Acoust. Soc. Am. 60, Suppl. 1, S76(A) (1976)]. Human speech was used as a control. A sign test was used to rank the systems; in both experiments the ordering was human speech, MISS, Allen/Klatt, Votrax. The pairwise comparisons were significant at less than the 1% level, except for three cases involving the control, MISS, and Allen/Klatt. A learning study was done for the letter experiment using linear probability models, whose results corroborated the ranking determined from the sign test. Study was made of individual problem sounds for each system in terms of human vocal parameters. [Supported by NSF grant SED74-15016-A02.]

2:42

BB7. Recognition of spoken spelled names applied to directory assistance. A.E. Rosenberg and C.E. Schmidt (Acoustics Research Department, Bell Laboratories, 600 Mountain Ave., Murray Hill, NJ 07974)

The automatic word recognition system originated by Itakura has been modified to accept strings of spelled letters spoken in isolation. The output of the recognizer is a set of best candidate letters for each letter spoken in the string. Candidate strings forming spelled last names and initials are compared to name strings extracted from a telephone directory stored in a disk file. A systematic search is carried out to find a matching entry in the directory. The outcome of a search is one or more matches or no match. An evaluation of the system was carried out using the 17 000 entry Bell Labs telephone directory. A list of 50 randomly selected names was extracted from the directory. Ten talkers participated in the evaluation spelling out each name in the list over a dialed-up telephone line input to the system. In the initial input operation, the median error rate in acoustic recognition of the individual spoken letters may be as high as 20%, but the constraints imposed by the system in making the directory search reduce the error to less than 6% in the final identification of the name.

2:54

BB8. Effects of various types of speaker normalization on an automatic vowel recognition scheme. S.A. Sroka (Department of Defense, Fort George G. Meade, MD 20755)

This paper compares the effect of non-normalization, uniform normalization, and formant-dependent normalization on an automatic vowel recognition scheme. The data analyzed is that of Peterson and Barney [J. Acoust. Soc. Am. 24, 175–184 (1952)]. Vowel identity is determined by observed formant values. Normalization is accomplished by a multiplicative transformation of a talker's formant data such that the talker's average formant frequency agrees with some reference value. The results indicate that normalization does significantly enhance vowel identification and that the linear effects account for most of the male/female differences under the model used.

3:06

BB9. Probabilistic vector model for voicing mode identification of intervocalic stop consonants. T.J. Edwards (Department of Speech and Hearing Sciences JI-05, University of Washington, Seattle, WA 98195)

A probabilistic vector model was developed to identify the voicing mode of phonemically equivalent intervocalic stop consonants with an accuracy commensurate with trained listeners, such performance to be attained independently of the succeeding vowel's identity or the differences among talkers. Acoustic features known to be salient to the perception and production of stop consonants were studied as random processes whose probability density functions provided models for voicing mode identification. Each acoustic feature present in a stop's production then contributed a vector whose magnitude and direction were determined from the probability distributions resulting from these density functions. Stop recognition was the result of summing these individual vectors, with the resultant vector specifying the voicing mode of the stop under consideration. Acoustic features incorporated into the model were closure duration, stop-consonant duration, fundamental frequency, adjacent vowel amplitudes, burst amplitude, formant transitions, voicing during closure, and voice onset time. When the performance of the probabilistic vector model was compared with that of trained listeners, correct voicing mode identification was approximately 99.0% in both cases.

3:18

BB10. Perplexity—a measure of the difficulty of speech recognition tasks. F. Jelinek, R.L. Mercer, L.R. Bahl, and J.K. Baker (Computer Sciences Department, IBM Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598)

Using counterexamples, we show that vocabulary size and static and dynamic branching factors are all inadequate as measures of speech recognition complexity of finite state grammars. Information theoretic arguments show that *perplexity* (the logarithm of which is the familiar entropy) is a more appropriate measure of equivalent choice. It too has certain weaknesses which we discuss. We show that perplexity can also be applied to languages having no obvious statistical description, since an entropy-maximizing probability assignment can be found for any finite-state grammar. Table I shows perplexity values for some well-known speech recognition tasks.

	Perplexity		Vocabulary	Dynamic
	Phone	Word	size	branching factor
IBM-Lasers	2.14	21.11	1000	1000
IBM-Raleigh	1.69	7.74	250	7.32
CMU-AIX05	1.52	6.41	1011	35