# Speaker-independent isolated word recognition using a new matching method

I. Nose, K. Mizuno and K. Yamada

Citation: The Journal of the Acoustical Society of America **71**, S7 (1982); doi: 10.1121/1.2019568

View online: https://doi.org/10.1121/1.2019568

View Table of Contents: https://asa.scitation.org/toc/jas/71/S1

Published by the Acoustical Society of America

JASA
THE JOURNAL OF THE ACOUSTICAL SOCIETY OF AMERICA
Special Issue:
Additive Manufacturing and Acoustics
Read Now!

MRP baseband system at 16.0 kb/s: 87.4 vs 93.9. Listening suggests that the speech quality from our approach is closer to the MRP quality for continuous speech than for the DRT words. For some applications, the computational efficiency and stability of our approach could outweigh the quality advantage of the MRP.

10:05

**C7. Properties of large lexicons: Implications for advanced isolated-word recognition systems.** Victor W. Zue and David W. Shipman (Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139)

A limitation of pattern-recognition-based, isolated-word recognition (IWR) systems is that computation and storage grows (essentially) linearly with the size of the vocabulary. When the size of the lexicon is very large (over 10 000 words), the computation and storage requirements associated with current IWR systems will become prohibitively expensive. Even if cost is not an issue, the performance of these IWR systems for a very large vocabulary is questionable. As part of our goal to design large-vocabulary, speaker-independent IWR systems, we investigated the properties of large lexicons ranging from 1250 words to 20 000 words. We focused on the question: Can properties of the lexicon be exploited in the design of the acoustic analyzer? Our results indicate that the search space can be substantially reduced without having to specify the *detailed* phonetic characteristics of the lexical entries. For example, if we hypothesize that the acoustic analyzer can only make a distinction between consonants and vowels, then the specification of the CV pattern of a given word can, on the average, prune the 20 000 word lexicon down to less than 1%. If, on the other hand, we hypothesize that the acoustic analyzer can make a six-way distinction, then roughly ⅓ of the lexicon has unique patterns. The effects of noise (i.e., the introduction of acoustic analyzer errors) and lexicon size will also be discussed. [Work supported by NSF.]

10:17

**C8. Speaker-independent isolated word recognition using a new matching method.** I. Nose, K. Mizuno, and K. Yamada (Oki Electric Industry Company, Ltd., 550-5 Higashi-asakawa-cho, Hachioji-shi, Tokyo 193, Japan)

This paper describes a speaker-independent isolated word recognition system which accepts telephone line speech. A recognition method is named selective weighted matching (SWM) which uses a weighted distance measure. The input speech signal is frequency-analyzed every 10 ms by a filter bank. The individual glottal characteristic is normalized frame by frame using a least-square-fit line of the speech spectrum. Each reference pattern has a specific region in the time-frequency domain. In the matching process of that region, the weighted distance computation is carried out under the predetermined condition. In the computer simulation of telephone line speech, we got the recognition accuracy greater than 96% with 12 words (digits and two command words in Japanese) spoken by 130 talkers. The same result was also obtained in the recognition test of the prototype machine.

10:29

**C9. Strategy for high quality speech elements formant synthesis.** P. Badin (Laboratoire de la Communication Parlée, E.N.S.E.R.G., 23 avenue des Martyrs, 38031 Grenoble Cedex, France)

To obtain high quality replication of natural speech elements using a formant synthesizer and investigate distinctive features of phonemes, we must determine the evolution of the 19 control parameters of the synthesizer. But there is no algorithm capable of automatically tracking all these parameters. Therefore we have developed a method of computer-aided manual synthesis. The analysis–synthesis process consists of three steps. The first one is an automatic analysis of the natural signal: analysis by linear prediction (to get poles and bandwidths) and pitch detection by the SIFT algorithm. Then the acquisition of certain parameters ($F0$, formants, bandwidths, and voicing amplitude) is manually carried out using the previous results. The third step is iterative: the other parameters are determined by the trial and error method; to perform the comparison between the original and synthesized signal, we use the FFT, a visual inspection of temporal signal and the "substitution" method (hearing the original sound and a sound composed of the original signal in which one part is replaced by the synthesized signal). This method may be long, but leads to high quality synthesis. We are now experimenting with this method in the synthesis of the French voiced fricatives ([v], [z], [ʒ]) in the context CVCV.

10:41

**C10. Analysis and synthesis of fundamental frequency contours of English sentences.** Hiroya Fujisaki and Keikichi Hirose (Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo, 113 Japan)

The voice fundamental frequency contour (henceforth the $F_0$ contour) has been known to be an important acoustical correlate of sentence intonation in English. Quantitative relationships between the characteristics of an observed $F_0$ contour and the underlying linguistic information, however, have not been fully clarified. In the present paper, we apply a method of analysis of $F_0$ contours originally proposed for the analysis of sentence intonation in Japanese. The method is based on a quantitative model of the process of conversion from discrete linguistic units into an observed $F_0$ contour, represented by the logarithm of voice fundamental frequency as a function of time. The validity of the model for English intonation has been demonstrated, firstly by analysis-by-synthesis, i.e., by showing that the model can produce close approximations to actual $F_0$ contours of English sentences, and secondly by perceptual experiments in which the naturalness of intonation was verified for synthetic sentences with $F_0$ contours generated by rule based on the proposed model. [Work supported by Ministry of Education Grant-in-Aid for Scientific Research No. 56580020 and No. 00551101.]

10:53

**C11. Length in Lingua.** Catherine P. Browman (Bell Laboratories, Room 2D-546, Murray Hill, NJ 07974)

Lingua is a program used for synthesizing speech from stored LPC-encoded demisyllables using rules to modify the duration, as well as other aspects, of the stored speech [C. P. Browman, ICASSP 80 Proceedings, 561–564 (1980)]. The original duration rules simply compressed nonphrase-final demisyllables a certain amount, depending on their stress level. The present duration rules, while still quite simple, are based on a different approach in which a variety of factors independently interact to produce the final durations of the subportions of the demisyllables. The factors include binary representation of stress (stress/unstress), binary representation of within-phrase position (final/nonfinal), and a variable factor based on the number of syllables in a set unit. The most recent output from Lingua, using the new rules, will be demonstrated.

11:05

**C12. Noise reduction in speech using adaptive filtering I: Signal processing algorithms.** R. W. Christiansen, D. M. Chabries, and D. Lynn (Electrical Engineering Department, Brigham Young University, Provo, UT 84602)

Adaptive filtering is employed in configurations to filter narrowband speech corrupted by noise. These configurations utilize either an independent sample of the input noise or rely on correlation properties of the speech to accomplish cancellation. Necessary constraints on the algorithms to retain and/or improve intelligibility for normal and hearing-impaired populations are presented. Previous work with