An acoustic-phonetic data base

William M. Fisher, Victor Zue, Jared Bernstein, et al.

Citation: The Journal of the Acoustical Society of America **81**, S92 (1987); doi: 10.1121/1.2034854
View online: https://doi.org/10.1121/1.2034854
View Table of Contents: https://asa.scitation.org/toc/jas/81/S1
Published by the Acoustical Society of America

of beat interval values of 6 notes, e.g., 313144. Musicians judged similarity of standard-comparison pattern pairs. On target trials, the comparison had the same rhythm as the standard; on related trials, the comparison had the standard's temporal contour, that is, the same series of longs, shorts, and sames without regard to interval size (as in the pair 313144--315133); on lure trials, the comparison had a different contour and intervals. Twenty-four musicians were randomly assigned to one of three listening conditions that differed in trial types presented: (A) targets versus lures; (B) targest versus relateds, or (C) relateds versus lures. Two other factors were tempo and metrical simplicity (i.e., the extent to which groups of intervals formed simple submultiples of the 16 beats). Discrimination was better (1) in condition A than in B or C, although performance in both B and C was still above chance, demonstrating the importance of interval and contour information, respectively; (2) when the comparison tempo was the same as the standard's; (3) with metrically simpler standards. Tempo interacted with condition: discrimination of comparison patterns that were faster or slower was poorer in B and C than in A.

11:20

NN7. Expressive microstructure in music: A first assessment of "composers' pulses." Bruno H. Repp (Haskins Laboratories, 270 Crown Street, New Haven, CT 06511-6695)

According to a provocative theory set forth by Manfred Clynes [most recently in Cognition and Communication **19**, No. 2 (1986)], there are composer-specific patterns of (unnotated) musical microstructure that, when discovered and realized by a performer, help to give the music its characteristic expressive quality. Clynes, relying on his own judgment as an experienced musician, has derived such "pulses" for several famous composers by imposing time-amplitude warping patterns on computer-sythesized classical music. To conduct a formal perceptual asessment of four such pulses, two sets of piano pieces by Beethoven, Haydn, Mozart, and Schubert, one in triple and the other in quadruple meter, each synthesized with each composer's pulse plus a "neutral" version, were obtained from Clynes and presented in random order to listeners of varying musical sophistication for preference judgments. The results show reliable changes in listeners' pulse preferences across different composers' pieces, which supports one essential prerequisite of Clynes' theory. However, there were some significant deviations from the predicted perference patterns. Possible causes will be discussed. [Work supported by NIH-BRSG.]

11:35

NN8. Effects of interpretation on timing in piano performance. Caroline Palmer (Psychology Department, Uris Hall, Cornell University, Ithaca, NY 14853)

Three timing methods in piano performance were described in a previous report [C. Palmer, J. Acoust. Soc. Am. Suppl. 1 **79**, S75 (1986)]: onset asynchronies, rubato patterns, and legato and staccato patterns. The present study examined the relation of these timing methods to performers' intended interpretations of an excerpt. Notated interpretations included indication of primary melody, phrasing, dynamics, and tempo changes. Performances were recorded on a computer-monitored (MIDI interface), velocity-sensitive, weighted keyboard. The performances showed consistent temporal patterns, directly related to the specified interpretations. Onset asynchronies occurred within chords, such that the primary melody (as notated by each performer) preceded the other voices. Asynchronies were generally largest on the first beat of each measure, marking the excerpt's metrical structure. Rubato patterns (deviations in tempo from mechanical regularity) showed larger changes between, than within performers' notated phrases. Legato and staccato patterns within phrases were accurately predicted by a combination of the durations of successive notes in the musical score, but only when performers' intended phrasings were taken into account. Each of these timing patterns decreased in degree when the pianists were asked to play unmusically. [Work supported by NSF and NIMH.]

FRIDAY MORNING, 15 MAY 1987 REGENCY BALLROOM A & B, 8:30 TO 11:45 A.M.

Session OO. Speech Communication IX: Speech Recognition

George D. Allen, Chairman
Department of Audiology and Speech Sciences, Purdue University, West Lafayette, Indiana 47907

Chairman's Introduction—8:30

Contributed Papers

8:35

OO1. An acoustic-phonetic data base. William M. Fisher (Texas Instruments, Inc., Dallas, TX 75266), Victor Zue (Massachusetts Institute of Technology, Cambridge, MA 02139), Jared Bernstein (SRI International, Menlo Park, CA 94025), and David S. Pallett (National Bureau of Standards, Gaithersburg, MD 20899)

DARPA has sponsored the design and collection of a large speech data base. Six hundred and thirty speakers read ten sentences each. Two sentences were constant for all speakers; the remaining eight sentences were selected from a set of 450 designed at MIT and 1890 selected at TI from text sources. The set of sentences is phonetically rich, balanced, and deep. Although all recordings were made in Dallas, we sampled as many varieties of American English as possible. Selection of volunteer speakers was based on their childhood locality to give a balanced representation of geographical origins. The subject population is adult; 70% male; young (63% in their twenties); well educated (78% with bachelor's degree); and predominantly white (96%). Recordings were made in a noise-reducing sound booth using a Sennheiser headset microphone and digitized at 20 kHz. A natural reading style was encouraged. The recordings are complete, and time-registered phonetic transcriptions are being added to the 6300 speech files at MIT. A version of the complete data base (16-kHz

sample rate, with acoustic-phonetic transcriptions—approximately 50 megabytes of data) will be made available to researchers through the National Bureau of Standards. [Work supported by DARPA.]

8:47

OO2. Phonetic labeling and acoustic correlates for building Japanese speech data base. Yoshinori Sagisaka, Shigeru Katagiri, and Kazuya Takeda (Advanced Telecommunications Research Institute International, Osaka, Japan)

Fine description of a large amount of speech signals is indispensable to acquire effective acoustic-phonetic rules in speech synthesis and recognition. To build a finely labeled speech data base, manual labeling is carried out using digital sound spectrograms and additional acoustic parameters that reflect power and spectral characteristics. Through labeler training, labeling items are modified several times to decrease the deviation of segment boundaries and to hasten labeling speed. As a result, not only the usual phonemic categories, but also finer phonetic events (e.g., closure, burst, and aspiration for plosive consonants) are labeled. Moreover, multiple segment boundaries (e.g., boundaries between vowels and following fricative consonants) and inseparable portions (e.g., aspiration followed by a devocalized vowel) are specially marked. To ensure labeling quality, reliability tests and error analysis are carried out. Labeling criteria using these results will be used for a large scale data base construction.

8:59

OO3. Evaluation of ASR front ends in speaker-dependent and speaker-independent recognition. Jean-Claude Junqua (Speech Technology Laboratory, 3888 State Street, Santa Barbara, CA 93105)

This paper extends previous experiments of Tsuga and Hermansky [J. Acoust. Soc. Am. Suppl. 1 **80**, S18 (1986)]. Those experiments evaluated the effect of spectral model order, in automatic speech recognition (ASR), using a small alpha-numeric data base. PLP (perceptually based linear predictive) and LP (linear predictive) analyses were compared, using a cepstral and RPS (root power sums) metric. Those experiments dealing with a bigger data base were validated (104 words, ten speakers). PLP RPS front end is compared with about ten other ASR front ends (LP cepstrum, LP RPS, critical band,...). Experiments were run at various different analysis model orders. Results of speaker-dependent ASR show that the low-dimensional PLP analysis is a good alternative to high-dimensional LP or filter bank analysis. The index-weighted metric improves the recognition accuracy, making the recognition results more uniform across the speakers. The speaker-independent experiments (which use the templates of one male and one female speaker as references) confirm the superior performance of the PLP method for extracting speaker–independent information. Most of the errors are on the consonants; some improvements on the recognition process are being investigated.

9:11

OO4. Speaker-independent vowel classification based on fundamental frequency and formant frequencies. James Hillenbrand and Robert T. Gayvert (RIT Research Corporation, Rochester Institute of Technology, Rochester, NY 14623-3435)

A multivariate distance measure (MVD) was tested on a data base consisting of hand-measured fundamental and formant frequency values from ten English vowels produced by 29 male talkers and 27 female talkers. The primary purpose of the study was to determine what set of parameters produced the best classification performance. Results included: (1) classification accuracies as high as 90% were obtained using exclusively internal information (i.e., only acoustic measurements of the unknown token); (2) classification accuracies as high as 95% were obtained when internal information was combined with information describing the talker (e.g., average formant frequencies); (3) parameter sets using absolute formant frequencies performed better than sets using formant ratios or log formant distances; (4) a very small decrement in performance (1.1%) was observed when MVD was tested under conditions of no overlap between the talkers used to train the system and the talkers used to test the system; and (5) MVD did not need to be trained separately on male and female talkers. [Work supported by the Air Force Systems Command, Rome Air Development Center, and the Air Force Office of Scientific Research, Contract No. F30602-85-C-0008.]

9:23

OO5. Predicting stress and syllable boundaries from segmental timing. Sven Anderson, Robert Port (Department of Linguistics, Indiana University, Bloomington, IN 47405), and Daniel Maki (Department of Mathematics, Indiana University, Bloomington, IN 47405)

Recent studies [Reilly and Port, J. Acoust. Soc. Am. Suppl. 1 **78**, S21 (1985)] show that timing measurements can be used to discriminate among items in a small vocabulary. Current efforts question the extent to which timing measurements can be used to augment basic segmental knowledge in a continuous, real-time speech recognition situation. These experiments examine a single template of segments (vowel–strong fricative–stop–vowel) each of which is embedded in different words in 112 different sentences for four English speakers. Syllable and word boundaries occur between each of the four segments. Discriminant analysis is used to predict the stress patterns, possible syllabifications, and word boundaries from timing measurements in the templates of a fourth male speaker after training on three others. Results indicate that within the template a linear combination of simple timing measurements can be used as predictors of lexical stress (95% correct). Word and syllable boundaries are much less reliably predicted (60%–75%) and apparently bear no simple linear relationship to timing measurements. [Research supported by NSF.]

9:35

OO6. Acoustic phonetic representations for continuous speech recognition: Networks versus lattices. Robert A. Brennan and Michael S. Phillips (Department of Computer Science, Carnegie–Mellon University, Pittsburgh, PA 15213)

A lattice has been used to represent the acoustic phonetic hypotheses in the Carnegie Mellon speech recognition system. This lattice was produced by four separate modules. Each module independently located and classified a set of target segments. A network representation is now being used that explicitly defines allowable paths through the phoneme hypotheses. The network has two advantages over the lattice representation. First, acoustic phonetic information can be used to join segment hypotheses. Second, a network allows the use of context-specific phonetic classification. Initially the network has been produced from the lattices using acoustic rules and broad classification information to connect segments, to adjust boundaries between segments, and to fill in gaps between the segments. Context-specific classification is now being developed for phones that are particularly sensitive to context effects. This network representation is being evaluated in the framework of the overall Carnegie–Mellon system. A comparison of the network representation with the lattice representation will also be presented. [Work supported by DARPA.]

9:47

OO7. Learning phonetic features using connectionist networks. Raymond L. Watrous and Lokendra Shastri (Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104)

A method for learning phonetic features from speech data using connectionist networks is described. A temporal flow model is introduced in which sampled speech data flow through a parallel network from input to output units. The network uses hidden units with recurrent links to capture spectral/temporal characteristics of phonetic features. A supervised