Automatic measurement of intonation

the system of optimal quality is investigated. Another objective of this study is to explore whether a sometimes difficult task of free number production required in magnitude estimations could be replaced by cross-modality matches using lines of various lengths produced by subjects on a computer screen. The main concern here was related to the unknown effects of the limited width of the computer screen on the magnitude estimation task. [This research was made possible by grant No. 2589 from the EEC Espirit SAM project.] [a] Also at Univ. of Iowa City, IA.

2:35–2:50

Break

2:50

7SP7. An unsupervised method for learning to track tongue position from an acoustic signal. John Hogden, Philip Rubin, and Elliot Saltzman (Haskins Labs., 270 Crown St., New Haven, CT 06511)

A procedure for learning to recover the relative positions of the articulators from speech signals is demonstrated. The algorithm learns without supervision, that is, it does not require information about which articulator configurations created the acoustic signals in the training set. The procedure consists of vector quantizing short time windows of a speech signal, then using multidimensional scaling to represent quantization codes that were temporally close in the encoded speech signal by nearby points in a *continuity map*. Since temporally close sounds must have been produced by similar articulator configurations, sounds which were produced by similar articulator positions should be represented close to each other in the continuity map. Using an articulatory speech synthesizer to produce acoustic signals from known articulator positions, relative articulator positions were estimated from synthesized acoustic signals and compared to the synthesizer's actual articulator positions. High rank-order correlations, ranging from 0.92 to 0.99, were found between the estimated and actual articulator positions. Reasonable estimates of relative articulator positions were made using 32 categories of sound, and the accuracy improved when more sound categories were used.

3:05

7SP8. Description of contextual factors affecting vowel duration. Jan P. H. van Santen (AT&T Bell Labs., 2D-452, 600 Mountain Ave., P.O. Box 636, Murray Hill, NJ 07974-0636)

As an initial phase of a project on duration models for predicting segmental durations from contextual factors, two natural speech databases produced by a male and a female speaker with more than 50 000 manually measured segmental durations were analyzed. This large quantity of data made it possible to perform a detailed analysis of the effects of several contextual factors, including lexical stress, word accent, the identities of adjacent segments, the syllabic structure of a word, and proximity to a syntactic boundary. Among the key results were the following. (1) The contextual factors accounted for up to 90% of the variance, and reduced the within-vowel standard deviation by a factor of 3. (2) There were complex interactions between factors, in particular between boundary proximity and postvocalic consonant identity and between lexical stress and syllabic word structure. (3) The effects of adjacent segments were reducible to the effects of voicing and manner of production; effects of place of articulation were negligible. (4) Proximity to a boundary should be measured in terms of syllabic and segmental position, not in terms of the sum of the intrinsic durations of segments between the target and the boundary. The results were compared with data reported by Crystal and House [J. Acoust. Soc. Am. **83**, 1551–1573 and 1574–1585 (1988)].

3:20

7SP9. Improved vowel recognition using the discrete cosine transform. D. J. Burr (Comput. Systems Res. Dept., Bellcore, 445 South St., Morristown, NJ 07960)

Some experiments are described that use a discrete cosine transform (DCT) to represent vowel spectra for classification by a neural network. The recognition accuracy of a neural network trained by back propagation was compared to that of a simple Gaussian classifier. Twelve English vowel categories from the TIMIT database were used {aa, ae, ah, ao, ax, eh, er, ih, ix, iy, uh, uw}. Training was done in a speaker-dependent manner using 152 male speakers from all geographic regions. Eight 16-ms DFT spectra were computed at 8-ms intervals about the center of each vowel. Sixteen DCT coefficients were derived from the 250- to 4250-Hz interval in each 8-kHz spectrum. Average DCT and delta DCT vectors over the eight frames were used as input features. Additional features included the first six peak frequencies of the DCT spectrum and two pitch parameters from the DCT coefficients. Best performance was obtained by the neural network with all 40 input features, resulting in 58.2% recognition accuracy. This compares favorably to a cochleagram representation using the same vowel classes [Muthusamy *et al.*, ICASSP-90]. The DCT also appears to require fewer coefficients than an equivalent cepstrum-based vowel classifier [Burr, ICASSP-92].

3:35

7SP10. Automatic measurement of intonation. Gerard W. G. Spaai, Arent Storm, and Dik J. Hermes (Inst. for Perception Res./IPO, P.O. Box 513, NL 5600 MB, Eindhoven, The Netherlands)

If speech intonation is only represented by an unprocessed series of pitch measurements, the interpretation can be hampered by three factors. First, because of the presence of unvoiced parts in an utterance, the continuously perceived pitch contour is disturbed by interruptions. Second, in many cases, speech is characterized by involuntary pitch perturbations that either cannot be heard at all, or do not contribute to the perception of intonation. Third, the perceptual meaning of a pitch movement depends upon its position within the syllable, in many cases with respect to the vowel onset. A correct interpretation of a pitch contour, therefore, requires the position of the vowel onsets to be known, too. These problems can be solved by interpolating the pitch at unvoiced parts from the adjacent pitch measurements, by removing all perceptually irrelevant details from the contour, and by indicating the vowel onsets. Besides presenting the procedures, a system will be presented which performs these tasks in real time, and which is currently used in an evaluation of its usefulness in teaching intonation to deaf persons. The applicability for the use of this intonation meter in training intonation of foreign languages will be indicated. [Work supported by Instituut voor Doven, St.-Michielsgestel, Netherlands.]