Statistical trajectory models for phonetic classification

William Goldenthal and James Glass

Citation: The Journal of the Acoustical Society of America **95**, 2876 (1994); doi: 10.1121/1.409413
View online: https://doi.org/10.1121/1.409413
View Table of Contents: https://asa.scitation.org/toc/jas/95/5
Published by the Acoustical Society of America

Angami is a Tibeto–Burman language spoken in the Naga Hills in the northeastern part of India. This study describes the phonetic inventory of one of the smaller Angami dialects, Khonoma, which is spoken by no more than 5000 people in the extreme west of the Angami region. Data for the study include recordings of two female and four male adult native speakers, along with palatographic samples and aerodynamic data for selected phonemes. Khonoma Angami has 45 consonants, 6 vowels, and 4 level tones. The plosives have a three-way contrast in voice onset time; nearly all other consonants have a two-way contrast, including the nasals and approximants. Angami voiceless nasals are different from the common type exemplified in, for example, Burmese. Aerodynamic measurements show that glottal opening during Angami voiceless nasals lasts until after the supralaryngeal articulatory closure is released, whereas in Burmese the glottis closes well before closure release. Thus Angami voiceless nasals could be described as "aspirated." The inaudibility of formant transitions between the nasal and the following vowel leads to interesting questions of how place of articulation is perceived in initial voiceless nasals. [Work supported by NSF.]

2pSP36. Temporal masking in automatic speech recognition. M. Pavel and Hynek Hermansky (Oregon Graduate Inst., P. O. Box 91000, Portland, OR 97291-10000)

It is reasonable to expect that relevant features for speech recognition should be the features that are well heard. Therefore the successful extraction of such relevant acoustic features must respect properties of the auditory periphery. To examine its temporal properties full masking (detection) and partial masking (loudness matching) experimental paradigms are used. The results of these masking experiments suggested a model consisting of a combination of linear and nonlinear components. Hence a recently introduced RelAtive SpecTrAl (RASTA) engineering technique for speech feature extraction [Hermansky *et al.*, Proc. Int. Conf. Acoustic, Speech, and Signal Process. (1993)] has been introduced, which employs linear temporal filtering of critical-band spectral energies done between two static nonlinearities and which has been successful as a feature extraction technique in automatic speech recognition. A correspondence between the results of the experiments and the behavior of the RASTA model has been found. Possible modifications of the RASTA technique to incorporate additional details of the experimental results will be discussed.

2pSP37. Phonetic structures of an endangered language: Montana Salish. Edward Flemming (Phon. Lab., Dept. of Linguist., UCLA, Los Angeles, CA 90024-1543)

Montana Salish, or Flathead, is an Interior Salishan language spoken by about 70 people on the Flathead reservation in Northwest Montana. This study utilized acoustic and aerodynamic data to examine three aspects of this language: (1) Like many other Salishan languages, Montana Salish permits extremely complex consonant clusters: initial sequences of five or more consonants are possible. (2) In addition to contrasts between plain and ejective stops, there are contrasts between glottalized and nonglottalized sonorants (lateral, nasals, glides, and pharyngeals). In Montana Salish, glottalized sonorants typically involve a glottal constriction early in the sonorant, i.e., they are pre-glottalized, rather than being produced with creaky voice throughout. (3) Montana Salish pharyngeals provide an interesting contrast to Semitic pharyngeals, in that they are very vocalic in character, sounding much like low back vowels. However, three distinguishing characteristics of these sounds have been identified: a pharyngeal constriction, resulting in *F*1 raising and *F*2 lowering, lowered fundamental frequency, and a change in voice quality. However there is considerable variability and not all of these properties are observable in all instances of these sounds.

2pSP38. Speech recognition with minimal spectral cues. Robert V. Shannon, Fan-Gang Zeng, John Wygonski, Vivek Kamath, and Micheal Ekelid (House Ear Inst., 2100 W. Third St., Los Angeles, CA 90057)

Speech recognition was measured in conditions that systematically reduced the amount of spectral information while preserving temporal envelope information. Speech stimuli were spectrally separated into several frequency bands. The temporal envelope in each band was extracted by half-wave rectification followed by low-pass filtering. Each resulting envelope was then used to modulate a noise band with the same bandwidth and cut-off frequencies as the original analysis band. Identification of consonants (16 consonants in aCa context), vowels (8 vowels in hVd context), and words in simple sentences (CUNY sentences) was measured as a function of the number and frequency distribution of analysis bands, the envelope filter cut-off frequency, and overall spectral shaping. Results as a function of the number of channels show that consonant recognition improves from one to two channels, but less improvement is observed from two to four channels. Vowel recognition improves significantly from one to three channels. Sentence recognition improves with the number of channels, approaching 100% correct with four channels. These results indicate that relatively little spectral detail is sufficient for recognition of speech. Results will be discussed in terms of speech processing strategies for cochlear implants. [Work supported by NIDCD.]

2pSP39. Phonetic structures of endangered languages: Observations and findings. Ian Maddieson and Peter Ladefoged (Phon. Lab., Linguist. Dept., UCLA, Los Angeles, CA 90024-1543)

A project to describe the phonetic structures of selected endangered languages was outlined in the preceding abstract. Under this project, data have been collected on a number of languages from East Africa, the Indian subcontinent, North America, Taiwan, and the South Pacific, and further fieldwork is planned. The analysis of these data continues to demonstrate that the languages of the world display greater diversity in their phonetic structures than many linguists anticipate. Illustrations will include Dahalo, a Cushitic language with lateral ejective affricates of unusual type and clicks that are invariably nasalized; Toda, a Dravidian language with an unusually rich inventory of consonants confined to code position, including three different coronal trills (each of which may also be palatalized); and Iaai, and Austronesian language with an explanded set of vowels and an unusually large set of voiceless sonorants. Without knowledge of the kind provided by our research, linguists and speech scientists would be in danger of assuming that the phonetic patterns known from familiar languages demonstrate the full range of sounds that can be utilized in possible human languages. [Work supported by the National Science Foundation.]

2pSP40. Statistical trajectory models for phonetic classification. William Goldenthal and James Glass (Spoken Language Systems Group, Lab. for Comput. Sci., MIT, 545 Technology Sq., Cambridge, MA 02139)

This talk presents phonetic models that capture both the dynamic characteristics and the statistical dependencies of acoustic attributes in a segment-based framework. The approach is based on the creation of a *track*, T_α, for each phonetic unit α. The track serves as a model of the dynamic trajectories of the acoustic attributes over the segment. The statistical framework for scoring incorporates the auto- and cross-correlation properties of the track error over time, within a segment. On a vowel classification task [W. Goldenthal and J. Glass, "Modeling Spectra Dynamics for Vowel Classification," *Proc. Eurospeech 93*, pp. 289–292, Berlin, Germany (1993)], this methodology achieved classification performance of 68.9%. This result compares favorably with other studies using the TIMIT corpus. This talk extends this result by presenting context-independent and context-dependent experiments for all the phones. Context-independent classification performance of 76.8% is demonstrated. The key to implementing the context-dependent classifier consists of merging tracks trained separately on left and right contexts to synthesize any desired context during classification. This method allows one to synthesize a track for triphone contexts not seen in the training set. Using a total of 4167 gender-dependent biphone tracks, 58 phonetic statistical models, and no phone grammar, a context-dependent classification performance of 80.5% was achieved. This result increases to 85.8% when a trigram phone grammar is added.

2pSP41. Phonetic structures of endangered languages: Investigative techniques. Peter Ladefoged and Ian Maddieson (Phon. Lab., Linguist. Dept., UCLA, Los Angeles, CA 90024-1543)

The survival of many languages is endangered as the speakers die