2:30

2pSCa5. Glottal articulations in tense vs lax phonation contrasts. Jianjing Kuang (Linguist, Univ. of Pennsylvania, UCLA Campbell Hall 3125, Los Angeles, California 90095, kuangjj@gmail.com) and Patricia Keating (UCLA, Los Angeles, CA)

This study explores the glottal articulations of one type of phonation contrast—the tense vs lax phonation contrasts of three Yi (Loloish) languages—which is interesting because neither phonation type is very different from modal voice, and both are independent of the languages' tonal contrasts. Electroglottographic (EGG) recordings were made in the field, and traditional EGG measures showed many small but significant differences between the phonations. Tense phonation involves more overall contact and briefer but slower changes in contact. Functional Data Analysis was then applied to entire EGG pulse shapes, and the resulting first principal component was found to be mostly strongly related to the phonation contrasts, and correlated with almost all the traditional EGG measures. Unlike the traditional measures, however, this component also captures differences in abruptness of contact. Furthermore, previously-collected perceptual responses from native speakers of one of the languages correlated better with this component than with any other EGG measure or any acoustic measure. The articulatory differences between these tense and lax phonations, involving glottal aperture and how glottal closure is made, are not extreme, but apparently they are consistent enough, and perceptually robust enough, to support this linguistic contrast. [Work supported by NSF.]

2:45–3:00 General Discussion

3:00–3:30 Break

3:30

2pSCa6. The effect of non-linear dimension reduction on Gabor filter bank feature space. Hitesh A. Gupta, Anirudh Raju, and Abeer Alwan (Elec. Eng., Univ. of California Los Angeles, 550 Veteran Ave., Apt. 102, Los Angeles, CA 90024, hiteshag@ucla.edu)

In this paper, we modify the Gabor feature extraction process, while applying the Gabor filters on the power-normalized spectrum and concatenating with power normalized cepstrum coefficients (PNCC), for noise robust large vocabulary continuous speech recognition. In Chang *et al*., ICASSP (2013), a similar Gabor filter bank (GBFB) feature set with multilayer perceptron (MLP) processing (to reduce the feature dimension) has been used with mel frequency cepstrum coefficients showing improvements on Aurora-2 and renoised Wall Street Journal corpora. On a subset of the Aurora-4 database (only male), our method has shown promising results (when using PCA) being 7.9% better than 39-dimensional PNCC features. But, the GBFB features are a rich representation of the speech spectrogram (as an overcomplete basis), and an appropriate dimension reduction/manifold learning technique is the key to generalizing these features for the large vocabulary task. Hence, we propose the use of Laplacian Eigenmaps to obtain a reduced manifold of 13 dimension (from a 564-dimensional GBFB feature set) for the training dataset with a MLP being used to learn the mapping so that the same can be applied to out-of-sample points, i.e., the test dataset. The reduced GBFB features are then concatenated with the 26-dimension PNCC plus acceleration coefficients. This technique should lead to better accuracies as speech lies on a non-linear manifold rather than a linear feature space. [This project was supported in part by DARPA.]

3:45

2pSCa7. Language material for English audiovisual speech recognition system development. Andrzej Czyzewski (Multimedia Systems Dept., Gdansk Univ. of Technol., Narutowicza 11/12, Gdansk 80-233, Poland, ac@pg.gda.pl), Tomasz Ciszewski, Dorota Majewicz (Faculty of Philology, Univ. of Gdansk, Gdansk, Poland), and Bozena Kostek (Multimedia Systems Dept., Gdansk Univ. of Technol., Gdansk, Poland)

The bi-modal speech recognition system requires a 2-sample language input for training and for testing algorithms which precisely depicts natural English speech. For the purposes of the audio-visual recordings, a training data base of 264 sentences (1730 words without repetitions; 5685 sounds) has been created. The language sample reflects vowel and consonant frequencies in natural speech. The recording material reflects both the lexical word frequencies and casual speech sound frequencies in the BNC corpus of approx. 100m words. The semantically and syntactically congruent sentences mirror the 100m-word corpus frequencies. The absolute deviation from source sound frequencies is.09% and individual vowel deviation is reduced to a level between 0.0006% (min.) and 0.009% (max.). The absolute consonant deviation is 0.006% and oscillates between 0.00002% (min.) and 0.012% (max.). Similar convergence is achieved in the language sample for testing algorithms (29 sentences; 599 sounds). The post-recording analysis involves the examination of particular articulatory settings which aid visual recognition as well as co-articulatory processes which may affect the acoustic characteristics of individual sounds. Results of bi-modal speech elements recognition employing the language material are included in the paper.

4:00

2pSCa8. Characteristics of automatic and human speech recognition processes. Mark VanDam (Speech & Hearing Sci., Washington State Univ., PO Box 1495, Spokane, WA 99202, mark.vandam@wsu.edu) and Noah H. Silbert (Commun. Sci. & Disord., Univ. of Cincinnati, Cincinnati, OH)

In a previous report [VanDam and Silbert (2013) *POMA***19**, 060006], we investigated performance of a commercially available automatic speech recognition (ASR) system [LENA Research Foundation, Boulder, CO] on acoustic recordings from family speech in naturalistic environments. We found that the ASR more accurately labeled children over adults and fathers over mothers, and human judge labels included substantial individual variation. The present work extends previous work by investigating the possible sources for both machine- and human labeling decisions. Classification tree models were fit to several acoustic variables for machine- and human labels of *CHILD, MOTHER*, and *FATHER*. Results suggest that (a) fundamental frequency (f_0) and duration measures influenced label assignment for both machine and human classifications, (b) the error of the fitted models is lower for the machine labeling procedure than for human judges, (c) machine- and human decision processes use the acoustic criteria (i.e., f_0 and duration) differently, and (d) f_0 is more important than duration for all labelers. Results may have implications for improving implementation and interpretation of ASR techniques, especially as they are useful for understanding child language applications and very large, naturalistic datasets that demand unsupervised ASR techniques.

4:15

2pSCa9. Crying for help: The Frye hearing and forensic acoustic analyses in State of Florida vs George Zimmerman. Al Yonovitz (The Univ. of Montana, Dept. of Communicative Sci. and Disord., Missoula, MT 59812, al.yonovitz@umontana.edu), Herbert Joe (Yonovitz and Joe, LLP, Irvine, CA), and Joshua Yonovitz (Yonovitz and Joe, LLP, Missoula, Montana)

Neighborhood watch volunteer George Zimmerman observed suspicious activity and called police. In minutes, he and Trayvon Martin were in an altercation when Mr. Zimmerman shot and killed Trayvon Martin and claimed self defense. The State's audio experts evaluated the 9-1-1 audio recordings to determine who spoke which background phrases, if any. A Frye hearing is where the court determines the reliability and admissibility of an expert's opinion by determining whether an expert's methodologies are generally accepted within the relevant scientific community. The judge in this case ruled that the State's two audio experts would not be permitted to testify during the trial. While the judge found the "aural perception and spectral analysis... are sufficiently established to have gained general acceptance within the scientific community," she took exception to how they were applied in this case. This case provided a venue and forum for opinions on voice analysis, methodologies, standards, and the quality required of evidentiary audio for determination of speaker identification and elimination. Positions taken by witnesses, a critique of the Frye Hearing and the scientific basis for witness and legal conclusions will be discussed.

4:30

2pSCa10. Intensity slopes as robust measure for distinguishing glottalic vs pulmonic stop initiation. Sven Grawunder (Dept. of Linguist, Max Planck Inst. for Evolutionary Anthropology, Deutscher Platz 6, Leipzig 04103, Germany, grawunder@eva.mpg.de)

A novel cross-linguistically robust measure is introduced for the linguistically relevant distinction of pulmonic vs glottalic (ejective) stops.