Prof-Life-Log: Production of Conversational Speech as a Function of Varying Environment

Hynek Boril, Ali Ziaei and John H.L. Hansen

{hynek, ali.ziaei, John.Hansen}@utdallas.edu

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.

LENA FOUNDATION Conference: April 28-30, 2013 Denver, Colorado USA

Introduction (01)

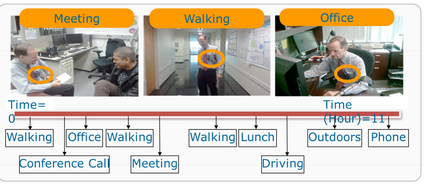
Objective

- Study speech production as a function of varying environment
- Majority of past literature – recordings in simulated/lab conditions (simulated noise/reverberation, lack of spontaneous communication)
- Prof-Life-Log corpus – acoustic footprint of daily activities of a university professor; natural conversational interactions and other acoustic events captured by LENA device

Outline

- Prof-Life-Log Collection
- Corpus Processing
- Data Analysis

Prof-Life-Log Collection (02)



- Unscripted speech collection in natural environments
- Unrestricted topics, vocabulary and language use

Corpus Processing (03)

Voice Activity Detection (VAD)

- Segments recordings into speech/non-speech segments

Environment Detection

- Acoustic signature vector (ASV) scheme identifies environment type from the non-speech segments

Primary/Secondary Speaker Detection

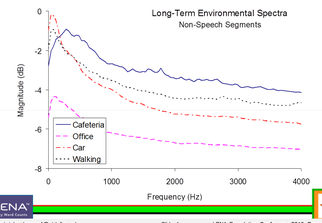
- Speaker identification system (SID) determines speaker identity

Phonetic Segmentation

- Automatic phone recognizer estimates phonetic content in speech segments and generates time labeled transcripts

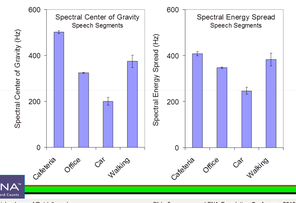
Environments: Long-Term Spectra (04)

- Total of ~12 hrs of audio used in this study
- Studied 4 environments: cafeteria, office, car, walking
- Each environment – characteristic spectrum (energy (dB), spectral slope…)



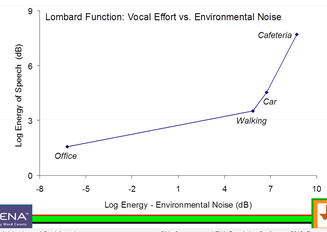
Spectral Center of Gravity (05)

- Spectral center of gravity (SCG) is sensitive to variations of speaking style and background noise (environment) and can be used for their reliable detection (plots with 95% conf. intervals)
- Spectral energy spread – rate of variance of spectral energy around SCG; complementary parameter to SCG in environment/talking style ID



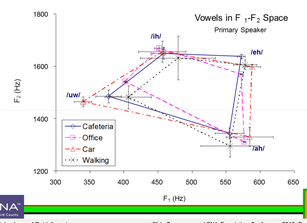
Lombard Function (06)

- Lombard function – dependency between the level of environmental noise and vocal effort. In general, individuals adjust their vocal effort to maintain intelligible communication over noise (Lombard effect)



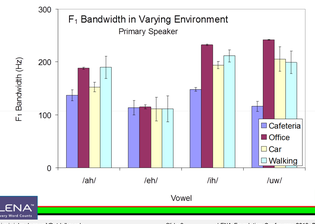
F_1-F_2 Formant Vowel Space (07)

- Automatic phone labels from phone recognizer combined with formant tracks from WaveSurfer
- Significant (95% confidence level) changes in vowel production across environments



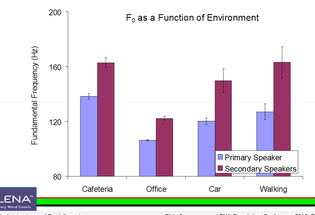
F_1 Bandwidth (08)

- Bandwidth of F_1 varies with type of environment
- For /ah/, /ih/ and /uw/ the environmental factor has significant effects



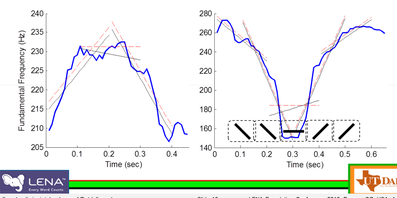
Fundamental Frequency F_0 (09)

- Mean fundamental frequency varies with environment.
- The primary speaker and the secondary speakers – the same trend in F0 changes → suggests the communicating parties use the same strategy to adjust to the environment or to each other

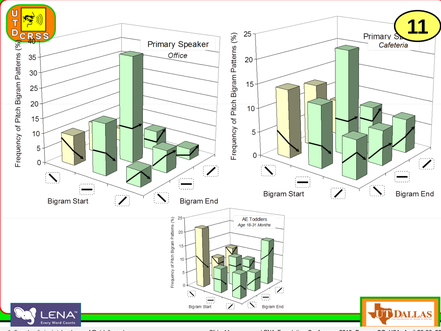


Pitch Patterns (10)

- Three pitch pattern elements (falling, flat, rising) matched to pitch contour
- Arbitrary parameters of the analysis – window length/step, thresholds to determine which patterns to still consider flat
- **Following slide** – bigram pattern histograms for office and cafeteria; note the prevailing flat pattern in office and increased pattern variability in cafeteria. Compare to 3rd plot – example of pitch patterns in children

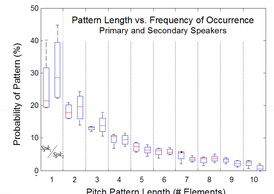


Primary Speaker Office / Cafeteria (11)



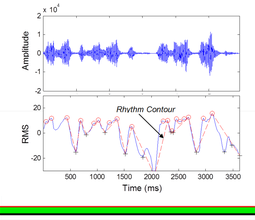
Pitch Pattern Duration (12)

- Pitch pattern duration is extracted per continuous island of voiced speech and can be used in analyzing the complexity of utterances.
- Note that the secondary speakers tend to produce shorter continuous pitch patterns than the primary speaker (a university professor)



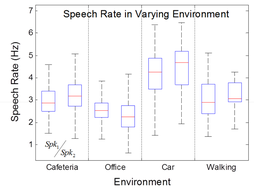
Speech Rate/Rhythm (I) (13)

- Speech rate (SR) is estimated from smoothed envelope of the speech signal
- Significant extremes in the envelope serve as candidates for SR detection
- Additional criteria are used to pick the most reliable candidates



Speech Rate/Rhythm (II) (14)

- Speech rate (SR) as a function of environment/communicating party
- Note that speech rate changes across environments and both primary and secondary speakers follow the same strategy in adjusting the rate



Conclusions (15)

- Introduced a set of automatic analyses in the naturalistic Prof-Life-Log corpus
- Demonstrated a number of speech production parameters that are sensitive to environment and/or the communicating party

Future Direction

- Extend the analyses to more Prof-Life-Log sessions (now ~12 hrs)
- Utilize keyword spotting in the analyses
- Develop a scheme to automatically assess emotions/level of stress in the recordings.