Acoustic Detectors for Conversational Interaction, Attitude and Affect

Automatic Speech Recognition (ASR) is the task of automatically producing text from speech. Extrapolations of the technological capabilities within the current paradigm have shown that computers will never reach human performance [1], which hints a need for a paradigm shift. But the original problem formulation: automatically producing text from speech is what you need for dictation, not for a dialog. Instead, let's formulate the problem as a Speaker is telling a story for a Listener. When humans does this, the Listener becomes an active listener, by producing short responses like "m", "mhm", "ja", "nja" and so on. These tokens are a subgroup of interjections, and serve many functions, were the most important is to neutrally signal that the listener hears that the speaker is talking. A listener response having this function is sometime called a back-channel or a continuer. They may also be used to signal interest/engagement [2] and attitude (news-receiving/dis-preference/neutral) [3] in general. The speaker who is telling the story has to elicit the responses and pay attention to the Listener, by picking up the signals produced by the listener and change the direction of the story accordingly. Typical adjustment are to tell more, to tell less, or to repeat/rephrase, all based on the short responses produced by the listener. The task for the attentive speaker has led to development of specialized detectors which are able to detect incoming speech as a listener response or not, given the timing constraints seen in human-human data [4]. The signaled attitude can then be further processed [3].

Figure 1: Interaction may be easier if one talk to computer as to a human.

When people speak, overwhelmingly one speaks at a time. Detecting incoming speech as listener response or not before while it is till produced, is a special case of this organizational turn-taking process. Listener responses are special since they are frequently interjected in overlap. In fact, around 41-45% of all speaker shifts occur after a minimally perceivable pause determined to be around 200 ms, while the rest occur before. To aid turn-taking in a dialog system, most research have focused on predictors based on prosodic (mostly the fundamental frequency) measurements for the speaker shifts which occur after a 200 ms pause, while research has been partly neglected for the speaker shifts which occur before a 200 ms pause. Predictors for these latter have been further developed based on the findings in [5]. Another approach is to model the joint interaction as a coupled Markov model, which is suitable for off-line processing of corpora [6].

The German linguist Ehlich states that since ancient times interjections have been considered an expression of mental state ("affectus animi"). Attitudes such as interest and amusement are cognitive states which may be viewed as emotions. The basic emotions are usually considered as

anger, disgust, fear, happiness, sadness, while examples of social/interpersonal emotions are affection, pride and shame. In [7], acoustic detectors for basic and social/interpersonal emotions are presented and evaluated on corpora recorded by actors. Acoustic detectors for a three class problem of negative/positive/neutral emotions, found in natural occurring human-computer and human-human dialogs, are presented in [8,9,10]. Detecting negative emotions in a dialog system may be useful as a sign of problems, i.e. the users becomes frustrated since he/she can't accomplish the task. If such emotions are detected, then the user may be forwarded to a call center. Analysis and acoustic detectors for natural occurring irritation resignation/neutrality and emotion intensity are presented in [11].

Processing large corpora for analysis of subtle variations in fundamental frequency, prosody and voice quality may be easier if explorative techniques are used. One may ask: How do I draw one fundamental frequency contour from many (See Figure 2)? How do I draw one spectrogram out of many (See Figure 3)? That is, how to visualize the "essence"?

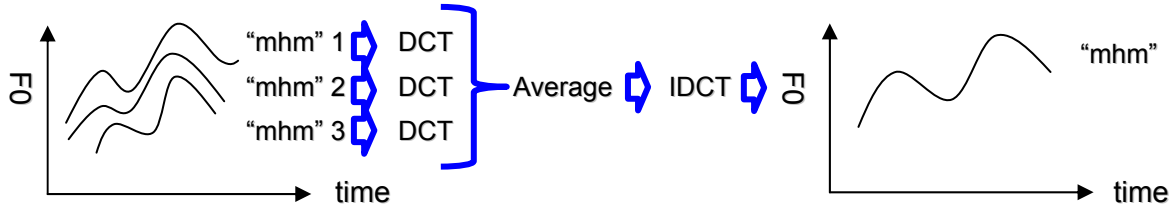

Figure 2: How to extract and visualize the prosodic "essence" via length-invariant discrete cosine transform. Two routes: 1) One-dimensional: given a pitch-tracker (shown in figure) 2) Two-dimensional: route given an average F0 estimate

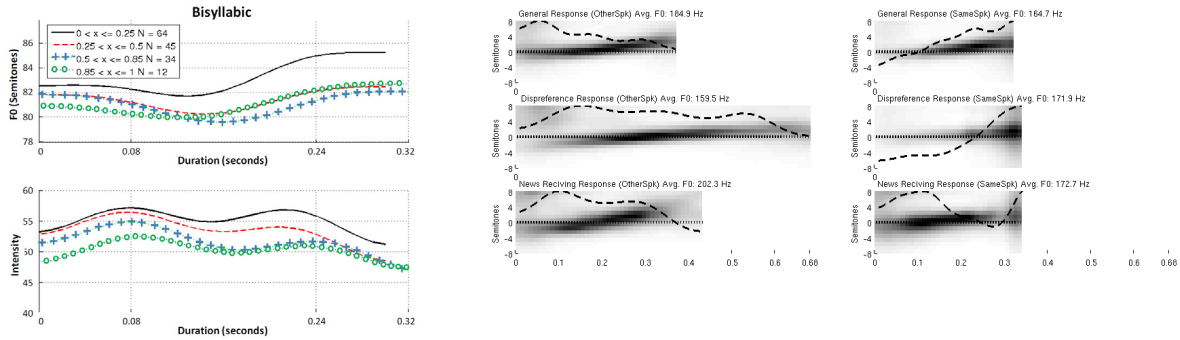

Figure 3: To the left (from [2]): Tappas Fogelbergs *"Mhm" as a function of relative position "x" for each call: Average F0 drops and curve becomes falter with respect to the second syllable. Average intensity drops and the peak of the second syllable drops. To the right (from [3]): F0 as densities relative to mean F0, Normalized Intensity as dashed lines. LDA analysis gives. Attitudes in backchannels can be classified based on prosody. Attitudes in isolated backchannels are easier to classify than turn-initial ones. Signaling attitude via prosody is not as important if you take the floor since there are more opportunities for signaling.*

References:

1. **Moore, R. K.** (2003) A comparison of the data requirements of automatic speech recognition systems and human listeners. In EUROSPEECH, Geneva, Switzerland, 2581–2584.
2. **Gustafson, J., & Neiberg, D.** (2010). Prosodic cues to engagement in non-lexical response tokens in Swedish. In *DiSS-LPSS*.
3. **Neiberg, D., & Gustafson, J.** (2010). The Prosody of Swedish Conversational Grunts. In *Interspeech, Special Session on Social Signals in Speech*.
4. **Neiberg, D., & Truong, K.** (in press). Online Detection Of Vocal Listener Responses With Maximum Latency Constraints. To be published in *ICASSP 2011*.
5. **Reidsma, D., Kok, I., Neiberg, D., Pammi, S., Straalen, B., Truong., K., Welbergen, H.** (submitted) Continuous Interaction with a Virtual Human. In Journal of Multimodal User Interaction.
6. **Neiberg, D., & Gustafson, J.** (2010). Modeling Conversational Interaction Using Coupled Markov Chains. In *DiSS-LPSS Joint Workshop 2010*.
7. **Neiberg, D., Laukka, P., & Ananthakrishnan, G.** (2010). Classification of Affective Speech using Normalized Time-Frequency Cepstra. In *Prosody 2010*.
8. **Neiberg, D., Elenius, K., & Burger, S.** (2009). Emotion Recognition. In Waibel, A., & Stiefelhagen, R. (Eds.), *Computers in the Human Interaction Loop* (pp. 96-105). Berlin/Heidelberg: Springer.
9. **Neiberg, D., & Elenius, K.** (2008). Automatic Recognition of Anger in Spontaneous Speech. In *Proc. of Interspeech 2008*. Brisbane.
10. **Neiberg, D., Elenius, K., & Laskowski, K.** (2006). Emotion Recognition in Spontaneous Speech Using GMMs. In *Proc. of Interspeech 2006*. Pittsburg.
11. **Laukka, P., Neiberg, D., Forsell, M., Karlsson, I., & Elenius, K.** (2011). Expression of Affect in Spontaneous Speech: Acoustic Correlates and Automatic Detection of Irritation and Resignation. *Computer Speech and Language, 25*(1), 84-104.