Paralinguistic speech processing

Paralinguistic speech processing (PSP) refers to analysis of speech signals with the aim of extracting information that is different from the linguistic content of speech (hence *para*linguistic = alongside linguistic content). Speaker recognition or verification is also traditionally considered as a separate problem, and does not fall within the scope of paralinguistic tasks.

The basic assumption is that the speech signal reflects the underlying cognitive and neurophysiological state of the speaker. This is since speaking involves highly complicated cognitive processing in terms of real-time communicative, linguistic, and articulatory planning. In addition, execution of these plans requires highly-precise motor control of articulators paired with real-time monitoring of the resulting signal, and both of these tasks take place in parallel with further speech planning. The overall physical characteristics of the speech production apparatus also shape the resulting signal by, e.g., changing the properties of the subglottal pressure and vocal tract filter. This means that any temporary or permanent perturbations in the cognitive and physical machinery may therefore show up in the speech of a person.

To give some examples, substantial cognitive load (e.g., a concurrent attention-requiring task) or neurodegenerative diseases affecting memory (e.g., Alzheimer's disease) may impact planning of the speech acts by reducing the efficiency of the involved cognitive resources, resulting in speech that differs from the typical speech from the same person in non-stressful or healthy conditons. In the same way, diseases affecting initiation of motor activities (e.g., Parkinson's disease) are likely to impact fluidity of speech production, and the symptoms will become more pronounced as the disease progresses. As for articulatory changes, stress and emotional distress can cause increased tension in the muscles of the larynx, which can result in tightening of the vocal folds and therefore also causes increases in the fundamental frequency of speech. Changes in the physical characteristics of the vocal tract may result from, e.g., having a cold. In this case, mucus on tract surfaces may affect resonance and damping characteristics of the vocal tract. In addition, the mucus may prevent full closing of the velum, causing nasalized speech often associated with a severe cold. Aging will also change the characteristics of the speech production apparatus, not only in childhood but also in later years of life (albeit more slowly than when the dimensions of the tract are still growing in the early childhood).

In addition to information that is not directly related to intended communicative goals, speech also contains paralinguistic characteristics related to communication. This is because speech has co-evolved with the development of other social skills in humans over thousands of years, and therefore speech (and gestures) can play different types of social coordinative roles beyond the literal linguistic message transmitted. For instance, prosody (or word choices) can reflect different social roles such as submissiveness or authority in different interactions. Attitudes and emotions showing up in speech can also be considered as communicative signals facilitating social interaction and cohesion, not just as speaker-internal states that inadvertently "leak out" for others to perceive. By demonstrating anger or happiness not just through visual (facial) gestures but also through voice, important information regarding social dynamics can be transmitted without requiring a constant visual contact between the interlocutors.

The basic aim of PSP is to use computational means to understand and characterize the ways that different non-linguistic factors shape the speech signal, and to build automatic systems for analyzing and detecting the paralinguistic factors from real speech captured in various settings.

Typical applications

Some typical applications of paralinguistic tasks include:

- Emotion detection
- Personality classification
- Sleepiness detection
- Analysis of cognitive or physical load
- Health-related analyses (cold, snoring, neurodegenerative diseases etc.)
- Speech addressee analysis (e.g., adult vs. child-directed speech)
- Age and gender recognition
- Sincerity analysis

Basic problem formulation and characteristics

The basic goal of paralinguistic analysis is to extract information of interest while ignoring the signal variability introduced by linguistic content, speaker identity, and other nuisance factors such as background noise and transmission channel characteristics. However, for some tasks, it may also be useful to analyse the language content of speech in order to infer information regarding the phenomena of interest.

Data sparsity

A relavely common property of PSP is that access to high-quality labeled data is limited. The data collection itself is often challenging and may include important ethical considerations, such as collecting data from intoxicated speakers or speakers with rare diseases.

Availability of good ground-truth labels for the speech data can also be difficult. For instance, judging of the underlying emotional states of speakers is difficult, whereas induced emotional speech by professional actors may not properly reflect the variability of emotional speech in real world communicative senarios. Assessing the severity of many diseases is also based on indirect diagnostics instead of having some type of oracle knowledge on a universally standardized scale. Every time humans are used for data labelling (e.g., assessing emotions), there is a certain degree of inter-annotator inconsistency due to differing opinions and general variability in human performance. This is the case even when domain experts are used for the task. Naturally, the more difficult the task or more ambiguous the phenomenon, the more there will be noise in the human-based ground-truth labels.

Finally, many types of interesting PSP data cannot be freely distributed to the research community due to data ownership and human participant privacy protection considerations. As a clear example, speech with metadata related to factors such as health or IQ of the speakers is highly sensitive in nature, and not all speakers consent to open distribution of their identifiable voice together with such private data of themselves. The data ownership considerations inherently limit the pooling of different speech corpora in order to build more comprehensive databases of speech related to different phenomena of interest, and generally slow down replicable open science. On the other hand, it is of utmost importance to respect the privacy of human participants in PSP (or any other) research—not only due to ethical considerations, but also since the entire field depends on access to data from voluntary human participants.

to be continued...

Computational Paralinguistic Challenge (http://www.compare.openaudio.eu/)

Research in paralinguistic speech processing has been largely advanced by an annual Computational Paralinguistic Challenge (ComParE) held in the context of ISCA Interspeech conferences. Every year since 2009, ComParE has included a number of paralinguistic analysis tasks with pre-defined datasets in which participants can compete with each other. Competitive baseline systems and results are provided to the participants as a starting point. New tasks and datasets can be proposed to challenge organizers, providing a useful channel for data owners and researchers to obtain competitive solutions to their analysis problems.

Further reading

Schuller, B. & Batliner, A. (2014). *Computational paralinguistics: Emotion, affect and personality in speech and language processing.* John Wiley & Sons Ltd, UK.