Annotating paralinguistic features in quasi-spontaneous speech

Adding the “vision” component?

Katarzyna Klessa & Maciej Karpiński
{katarzyna.klessa, maciej.karpinski} @amu.edu.pl

Introduction
In this presentation a planned future extension to an on-going project is to be discussed. The project is focused on identifying, describing and annotating paralinguistic features in quasi-spontaneous speech using audio data for speaker characterisation. The extension would rely on bringing the “vision” component to the scope of analysis and including it into the present framework, i.e. including features obtained from video recordings.

Background
Paralinguistic or non-linguistic features might be categorised or grouped in multiple ways depending on the prospective application and even their definition (although long discussed) is still not quite clear. Different categorisations might be expected from linguists, phoneticians, speech technologists or engineers, and further differentiation may result from distinguishing various levels of analysis. In some cases, it appears sufficient and adequate to distinguish a limited number of inherently diversified categories. However, for the needs of automatic recognition of informal, interactive speech as well as for speaker characterisation, a more sophisticated approach is required. The acoustic, phonetic or perceptual correlates of paralinguistic features have been recognised to a different extent, as well as their multi-lateral interactions and influences. Human perception of individual characteristics of speech assumes treating the whole range of co-occurring speech and non-speech events in a holistic way. Although various types of information might be processed separately, they interact during speaker or person recognition. Not only is it advisable to treat voices as multidimensional entities but also to consider at least some information related to the interactive character of speech communication as well as information about the environment and situational context which may significantly influence a speaker’s vocal behaviour. Consequently, it appears justified to adopt a wider range of cues to rely on in order to find the closest possible approximation of the real interpretations by human listeners.

Even in early, pioneering works on paralanguage, gestures were considered as a paralinguistic component of communication. Currently, there is a growing body of evidence that gestures are actually very close to language or that at least they can be efficiently analysed using the methods and tools of linguistics. As an almost indispensable component of utterances, gestures (as well as facial expressions and body positions) can be hardly excluded from the paralinguistic analysis.

Establishing unambiguous boundaries between features and their categories in the process of creating specifications and then the actual annotation of spontaneous or affective speech is a problematic task even for experts. With some features, it proved useful to use graphic feature continuum or space representation instead of arbitrary assignment of categories or parameters (cf. Queen’s University Belfast by Roddy Cowie et al. for the analysis of emotional speech using a two-dimensional space representation). A similar issue occurs in the categorisation and annotation of gestures; it has been suggested that traditionally acclaimed categories should be understood as forming a continuum.

Annotation System - the use of graphical representation of feature space
A graphic control enabling annotation of non-categorial / continuous features on a two-dimensional space users may create their own pictures as background styles for custom images, the program stores the clicked cursor position as two coordinates on the selected annotation tier.

The features might be defined by the user, e.g. affective states, voice quality, prosody perception, gesture, facial expression, body position. Aims of the use of graphical representation of the feature space:
• enabling annotation of continuous features with fuzzy categories
• a potential means of discovering new clusters of feature values, and subsequently defining new categories

Adding the “vision” component?
Further software development with a view to enable annotation of both audio and video recordings
Graphical representation of the feature space appears as very convenient in gesture annotation where a number of gradual dimensions should be taken into account simultaneously, and multiple functions of gesture should be represented in a possibly intuitive way.

Preliminary areas of interest:
• the perceptual judgement of facial expression (e.g. using a two-dimensional emotion feature space, see the picture above)
• physical properties of movements (e.g. two-dimensional space for the range and speed of realization)
• the gestural space (annotation of the gesture location in space).

Some other current features of the software & availability
• Multi-tier annotation
• XML annotation file format & import/export annotation from/to Transcriber, Wavesurfer, and TXT files
• Plugins in C# - a plugin architecture is built in the system providing easy access to the application methods and properties enabling flexible control of the program functions by C# scripts; the scripts may use external dll libraries
• Annotation System - available on-line in 2013 (free for research)

Institute of Linguistics, Adam Mickiewicz University, Poznań, Poland
Adam Mickiewicz University Foundation, Poznań, Poland

Acknowledgement:
This work is supported from the financial resources for science in the years 2010–2012 as a development project (project no. O R00 0170 12).