Online Diarization Enhanced by recent Speaker identification and Structured prediction Approaches

Résumé du projet de recherche (Langue 1)

Speaker diarization is an unsupervised process which aims to identify each speaker within an audio stream and to determine when each speaker is active. It considers that the number of speakers, their identities and their speech turns are all unknown. Speaker diarization has become an important key technology in many domains such as content-based information retrieval, voice biometrics, forensics or social-behavioural analysis. Example applications of speaker diarization include speech and speaker indexing, speaker recognition (in the presence of multiple speakers), speaker role detection, speech-to-text transcription, speech-to-speech translation and audiovisual content structuring. Although speaker diarization has been studied for almost two decades, current state-of-the-art systems suffer from many limitations. Such systems are extremely domain-dependent. For instance, a speaker diarization system trained on radio/TV broadcast news experiences drastically degraded performance when tested on a different type of recordings such as radio/TV debates, meetings, lectures, conversational telephone speech or conversational voice-over-IP speech. Overlapping speech, the spontaneous speaking style, background noise, music and other non-speech sources (laugh, applause, etc.) are all nuisance factors which badly affect the reliability of speaker diarization. Furthermore, most existing work addresses the problem of offline speaker diarization: the system has full access to the entire audio recording beforehand and no real time processing is required. Therefore, the multi-pass processing over the same data is feasible and a bunch of elegant machine learning tools can be used. Nevertheless, these compromises are not admissible in real-time applications mainly when it comes to public security and fight against terrorism and cyber-criminality. Moreover, after an initial step of segmentation into speech turns, most approaches address speaker diarization as a bag-of-speech-turns clustering problem and do not take into account the inherent temporal structure of interactions between speakers. Better performance may be achieved by integrating this information by exploiting structured prediction techniques to improve over standard hierarchical clustering methods. Speaker diarization is inherently related to speaker recognition. In recent years, the performance of state-of-the-art speaker recognition systems has improved enormously on account of new recognition paradigms such as i-vectors and deep learning, new session compensation techniques such as probabilistic linear discriminant analysis, and new score normalization techniques such as adaptive symmetric score normalization. However, existing speaker diarization systems do not take full advantages of these new techniques.

Résumé du projet de recherche (Langue 2)

The PhD programme will develop new technologies to overcome the limitations and weaknesses in the current state of the art of speaker diarization. The work will improve diarization performance through the application of the most recent automatic speaker recognition developments to speaker diarization, will facilitate on-the-fly processing with new online learning approaches and will pioneer entirely new approaches to speaker diarization which learn and exploit speaker turn models through structured prediction. Many security applications require online approaches to speaker diarization. This need arises chiefly from the processing of ‘big-data’. With the vast majority of work in speaker diarization being offline, the PhD will bring about a step change in the current research focus. New, online learning algorithms will be investigated in order to process vast datasets with manageable demands on both memory and computational processing capacity. Most existing approaches only take simple and local structure into account and completely overlook long-term and higher-order structure. This work will address speaker identification as a sequence labeling task and will use structured prediction techniques to account for the inherent temporal structure of interactions between speakers. While current strategies have been applied in multiple scenarios, e.g. for telephony speech, broadcast news, media collections and meetings, it is arguably the latter which presents the greatest challenge and the most relevant to the security scenario in this thesis. Meetings are unorchestrated, spontaneous exchanges involving varying numbers of speakers, acoustic and recording conditions and a high potential for overlapping speech, which is currently a bottleneck to performance. While there is some work in the literature which has investigated approaches to overlap detection and handling, it remains very much an unsolved problem. Also, while considerable progress has been made in the related research field of speaker recognition, relatively few recent developments have been explored in the context of speaker diarization.