Proposition de recherche doctorale

Normalisation linguistique pour la segmentation et regroupement en locuteurs et pour la reconnaissance de locuteur

Mots clés :

- Directeur de thèse : nicholas EVANS
- Co-encadrant(s) :
- Unité de recherche : Laboratoire de recherche d'EURECOM
- Ecole doctorale : École Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal: Divers

Résumé du projet de recherche (Langue 1)

Many automatic speech processing algorithms make numerous, sometimes unrealistic or restrictive assumptions about the use-case scenario. Some algorithms require low levels of noise, low reverberation, a particular grammar, or the presence of a single speaker, for example. This proposal relates to the handling of speech from multiple speakers, a problem referred to as speaker diarization. Otherwise known as the ‘who spoke when?’ problem, speaker diarization involves the detection of speaker turns within an audio document (segmentation) and the grouping together of all same-speaker segments (clustering). Recent work shows how the performance of state-of-the-art algorithms can be degraded by nuisance variation, e.g. linguistic influences. This thesis will develop new normalization approaches to attenuate linguistic influences and improve speaker diarization performance. The work will also consider the potential of linguistic normalization in the field of short-interval speaker recognition where performance can be degraded by the linguistic imbalance between training and testing data. Finally, the work will be conducted alongside Master’s projects to investigate prototype algorithms for smart phone mobile platforms including the convergence of speaker diarization and recognition with source separation and new applications in distributed, collaborative and real-time speech processing.

Résumé du projet de recherche (Langue 2)
Outline and objectives ===================== Speaker diarization is generally studied in an unsupervised scenario, in which there is no a priori knowledge regarding the speakers, or even the number of speakers. In consequence, today’s state-of-the-art approaches to speaker diarization generally involve an iterative search for a speaker inventory and segmentation using either agglomerative or divisive hierarchical clustering. As we showed in recent work [1], without any a priori knowledge of the speakers, both approaches have the potential to be affected by unwanted variation and can converge towards nuisance factors, rather than speakers. In the case of speaker diarization, and indeed text-independent speaker recognition, one of the greatest unwanted influences on a speech signal is that of linguistic content, i.e. the spoken text. In our most recent work [2] we showed that significant improvements in speaker diarization can be achieved through linguistic normalization. The new approach is referred to as phone adaptive training (PAT). It is used to learn a feature level transform which attenuates phone influences in the feature space while preserving differences between speakers. Early work shows it can deliver up to 33% relative improvement in the diarization error rate. The algorithm is based on constrained maximum likelihood linear regression (cMLLR) which allows for a transform to be learned for a single phone or a group of similar phones, depending on the quantity of available data. As such, the algorithm does not necessarily depend on entirely accurate, phone-level transcriptions. Our preliminary work in [2] assessed the performance of the algorithm in order to prove the concept with ground-truth word-level and speaker-level transcriptions. This PhD thesis will develop the idea to use a priori knowledge of speech at the phone level, e.g. from automatic speech recognition, so that it is entirely practical in nature and uses no ground-truth information. The work will also study the use of data from both the speech signal of interest and cMLLR transforms computed on external speech data. We will also be interested to see how the use of prior knowledge of the speakers can best be harnessed within an otherwise typical speaker diarization scenario. The second component of the proposed work will investigate factor analysis based approaches to speaker diarization. Factor analysis [3] and iVectors [4] have produced some of the greatest advances in the state of the art in the field of speaker recognition. We will investigate the parallels between intersession variation for speaker recognition and linguistic normalization for speaker diarization. The third component of the PhD thesis will assess the application of linguistic normalization to short-duration speaker recognition and in combined applications of speaker diarization and speaker recognition. Assuming text-independent speaker recognition, where models are trained on phonetically balanced speech, performance is often significantly degraded when the test segment is phonetically imbalanced, or of short duration. We will investigate the use of PAT to phone-normalize test segments and thus to improve recognition performance. This work will be conducted in speaker diarization / speaker indexing and conventional speaker verification scenarios. The final component involves the reduction of computational complexity and the investigation of new applications in distributed, collaborative and real-time speech processing. This work will be conducted both by the new PhD recruit and by Master’s projects. The current state-of-the-art approaches to speaker diarization are computationally prohibitive. We will investigate new work in the use of binary feature representations and modeling in the field of speaker recognition and their application in computationally efficient linguistic normalization and speaker diarization. This part of the work will lay the foundations for future real-time, computationally efficient implementations on mobile platforms. The detection and appropriate handling of source-sparse and overlapping speaker intervals is at the limits of the current state-of-the-art in both research fields. Source separation is often tackled blindly, i.e. with no prior knowledge of the speakers, while this is exactly the target of speaker diarization. The performance of speaker diarization is degraded by the presence of overlapping speech, while this can be detected and handled with source separation. The convergence of these two technologies, and further efforts to utilize speech acquired from multiple, spatially separated microphones, i.e. in a mobile device scenario, will lead to new applications in distributed, collaborative approaches to speech processing. The drive to improve computational efficiency will be paramount in supporting their implementation on low-power, portable devices. One Master’s project will investigate the convergence of speaker diarization, speaker recognition and source separation and whereas another will develop prototype algorithms for smart phone, mobile platforms. [1] N. Evans, S. Bozonnet, D. Wang, C. Fredouille and R. Troncy, “A comparative study of bottom-up and top-down approaches to speaker diarization,” IEEE Transactions On Audio Speech and Language Processing” (TASLP), special issue on New Frontiers in Rich Transcription, February 2012, Volume 20, N°2, ISSN: 1558-7916 [2] S. Bozonnet, R. Vipperla, and N. Evans, “Phone adaptive training for speaker diarization,” INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, September 9-13, Portland, Oregon, USA [3] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel, “Joint factor analysis versus eigenchannels in speaker recognition,” IEEE Transactions on Audio, Speech and Language Processing 15 (4), pp. 1435-1447, May 2007. [4] N. Dehak, “Discriminative and Generative Approaches for Long- and Short-Term Speaker Characteristics Modeling: Application to Speaker Veri?cation,” Ph.D. thesis, Ecole de Technologie Supérieure, Montreal, 2009.