Une meilleure classification des trames pour une meilleure reconnaissance du locuteur

Mots clés : reconnaissance du locuteur, réseau de neurones profonds, débruitage, classification

Des études récentes ainsi que notre propre expérience sur le bruit montrent qu'un bon système de reconnaissance du locuteur doit se reposer sur un **bon** classificateur de trames. D'ailleurs, en observant les formules qui permettent d'obtenir les i-vectors (approche état de l'art) [1], nous constatons que celles-ci sont fondamentalement basées sur l'association trame/classe. Un **bon** classificateur doit avoir deux caractéristiques importantes :

- Une classification aussi fine que possible (nombre important de classes)
- Doit faire le moins d'erreurs possible quand il associe les trames aux différentes classes

Il est bien évidemment possible de segmenter la parole en utilisant un système de reconnaissance de la parole. Dans ce cas , les classes sont les états des HMM utilisés par le système de reconnaissance
[2]. Mais cette solution impose le développement d'un système de reconnaissance (pour la langue en
question), ce qui est une tâche difficile et coûteuse comparé au coût de développement des systèmes de reconnaissance du locuteur actuellement état de l'art (à base de i-vector et de GMM-UBM).

Le candidat à cette thèse doit proposer des systèmes de classification de trames performants et robustes aux éventuelles distorsions (bruit). Le candidat explorera dans un premier temps deux pistes:

- La première consiste en l'utilisation des probabilités a posteriori issues du GMM-UBM, en effet ce vecteurs de probabilité donne un point de vue très intéressant sur la trame. Ces derniers mois, des études préliminaires nous ont incités à penser qu'une classification fondée sur ces vecteurs de probabilités pourrait largement dépasser celle fondée sur les trames elle-mêmes (une classe correspond une gaussienne dans le GMM-UBM).

- La deuxième piste consiste à utiliser les réseaux de neurones profonds afin de réaliser d'une manière non supervisée (ou semi-supervisée) une classification des trames de parole. Des centaines de millions de trames seront utilisées pour cela. Ces dernières années les réseaux de neurones profonds ont montré un succès indéniable pour ce type d'applications.

Le deuxième volet de cette thèse consiste à considérer les probabilités a posteriori comme des observations que nous pouvons représenter par un modèle statistique. Ces observations peuvent être distordues pour plusieurs raisons : bruits additifs ou bruits convolutifs ou autres. L'objectif sera de proposer des systèmes de débruitage de ces observations : il s'agit d'une approche très originale pour améliorer les performances des systèmes de reconnaissance du locuteur et aussi pour compenser les distorsions qui peuvent être dues aux différents types de bruits. Nous avons une bonne expérience sur le débruitage des trames [3] que nous souhaitons appliquer aux vecteurs de probabilités a posteriori.

[1] P.-M. Bousquet, D. Matrouf, J.-F. Bonastre, “Intersession compensation and scoring methods in the i-vectors space for speaker recognition”, in Proc. Interspeech, 2011.

[2] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, “Deep neural networks for extracting Baum- Welch statistics for speaker recognition,” in Proc. ICASSP, 2014.

[3] Waad Ben Kheder, Driss Matrouf, Pierre-Michel Bousquet, Jean-François Bonastre and Moez Ajili. “Robust speaker recognition using MAP estimation of additive noise in i-vectors space”,2 nd International Conference on Statistical Language and Speech Processing, SLSP 2014, Grenoble, France, October 14-16, 2014.