Large-scale prosodic-acoustic analysis of ASR errors in French

Camille Dutrey [1,2], *Martine Adda-Decker* [1,2], *Fabián Santiago* [1]

[1] LPP (UMR 7018) – Université Sorbonne Nouvelle & CNRS
[2] LIMSI-CNRS

{camille.dutrey,madda}@limsi.fr, fabian.santiago.ling@gmail.com

Abstract

Human speech recognition still outperform Automatic speech recognition (ASR). Despite the fact that ASR systems still produce errors (Word Error Rate around 30%), researches which diagnose them are marginal. One approach applied in such researches is to explore acoustic-prosodic characteristics of erroneous segments (phones or words) [1, 2], but deeper analysis is still needed to address this problem.

In the frame of the VERA ANR project, the main goal of this study is to produce a typology of ASR errors based on acoustic-prosodic characteristics of errors segments compared to non erroneous ones.

This exploratory study is conducted on the French ETAPE corpus [3], a 30h radio and TV broadcasts corpus, along the lines of the ASR error analysis that [4] did at a morpho-syntactic level on the same corpus. We use a transcription of reference provided by the ETAPE project while the ASR output (hypothesis) is provided by the LIUM laboratory.

The acoustic-prosodic analysis is performed at the phonetic and word levels. To do so, we conduce an automatic alignment on phones and words between the audio signal and the reference by using the LIMSI system [5]. With the same system, we re-align the ASR hypothesis. From these alignments and an automatic acoustic extraction made with `Praat` on the speech signal, we automatically calculate acoustic-prosodic features for phones and words (e.g. duration, f0, formants, *etc.*).

Since not all ASR errors are of equal importance, we discuss the acoustic-prosodic characteristics of phones and words – and their statistical relevance – according to different types of ASR errors, at the word level. In the first hand, we make a difference between three types of errors: word insertion, word deletion and word substitution. In the other hand, we classify errors with an edit distance between the reference and the hypothesis with the aim to isolate inflection errors (e.g. "cat" *vs* "cats") from "more serious" ones like homophones errors.

References

[1] J. Hirschberg, D. Litman, and M. Swerts, "Generalizing prosodic prediction of speech recognition errors," in *Proceedings of ICSLP*, 2000.

[2] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[3] G. Gravier, G. Adda, N. Paulsson, M. Carré, A. Giraudel, and O. Galibert, "The ETAPE corpus for the evaluation of speech-based TV content processing in the french language," in *Proceedings of LREC*, 2012, pp. 114–118.

[4] M. Goryainova, C. Grouin, S. Rosset, and I. Vasilescu, "Morpho-syntactic study of errors from speech recognition system," in *Proceedings of LREC*, 2014, pp. 3050–3056.

[5] J. Gauvain, L. Lamel, H. Schwenk, G. Adda, L. Chen, and F. Lefèvre, "Conversational telephone speech recognition," in *Proceedings of ICASSP*, 2003, pp. 212–215.