PhD Open Days

Developing speech therapy games for children with speech disorder

Training Network on **A**utomatic Processing of **PA**thological **S**peech (**TAPAS**)

Thomas Rolland (thomas.rolland@l2f.inesc-id.pt)

Introduction

Although adult automatic speech recognition (ASR) systems achieve a good recognition performance, there is a gap in terms of performance when the same automatic speech recognition systems are used on children (from two to five times worse for children speech than for adult speech). This is due to the high variability in children's speech and the lack of large amounts of available data for children. This high variability is present for children of the same age, between the same children at different ages and also for the same child at the same age. It is caused by physical and developmental changes in the vocal tract, that lead to spectral and temporal variability in the acoustic signal, but also by a partial linguistic and grammatical knowledge which that lead to mispronunciations of certain patterns and grammatical mistakes [1].

The first step to build a therapy game for children with a speech disorder is to define a good ASR baseline with healthy children. Towards this goal, I focus my work on two important points: (1) Adapt acoustic models from adult to children speech characteristics and (2) incorporate new relevant speaker features, that can provide complementary information that compensate for the acoustic differences of most conventional features

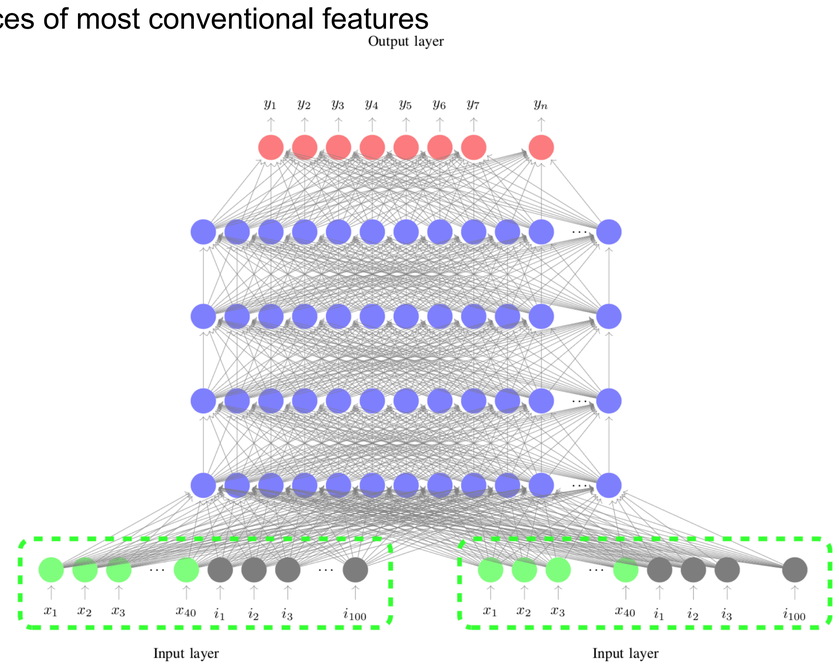

Figure 1: Acoustic Variability Modeling transfer learning (Red:Output, Blue: Hidden, Grey:I-vector, Green: MFCCs) [2]

Adapt acoustic model from adult to children using transfer learning

One way explored to solve this variability and lack of data for children speech is to adapt an adult acoustic model to children acoustic characteristics using transfer learning methods [2,3] (see Figure 1). My first baseline results on Portuguese data (see Table 1) show that transfer learning methods are relevant to solve these issues.

	Adult	children
Train with adult data	7.28	102.88
Train with children data	83.92	65.61
Transfer learning approach	26.01	64.91

Table 1: ASR baseline performance (Word Error Rate (%)) using MFCCs as input

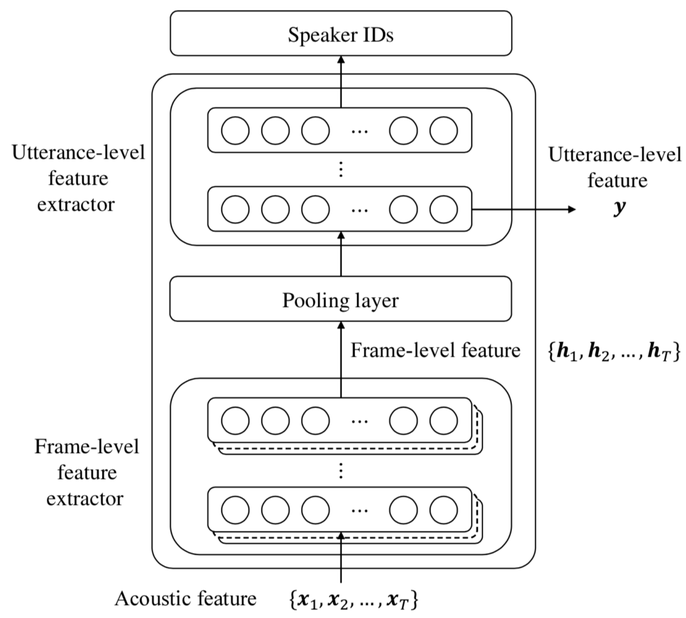

Figure 2: DNNs for extracting utterance-level speaker features (X-vector) [4]

Learning new relevant features

One way to have a robust acoustic model is to concatenate speaker information with acoustics information (MFCCs) , traditionally these speaker informations are I-vector (a sort of speaker voice-print that is estimated from each utterance). Similarly to this approach, I want to create a new speaker representation well adapted for children. I based my work on X-vector, a neural network alternative of I-vector. X-vector architecture has three different components (see Figure 2):

1- Frame-level features

2 - Pooling layer

3 -Utterance-level features

Traditionally, as pooling layer community use statistic pooling (mean and second-order derivation are computed and concatenated). In order to improve this pooling [4] provided an attention extension of this statistic pooling.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 766287.

This work was supported by national funds through Fundação para a Ciência e a Tecnologia (FCT) with reference UID/CEC/50021/2019.

References

[1] J. Kennedy, S. Lemaignan, C. Montassier, P. Lavalade, B. Irfan, F. Papadopoulos, E. Senft and T. Belpaeme, "Child speech recognition in human-robot interaction: evaluations and recommendations," in Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction. ACM Press, 2017, pp. 82–90.

[2] P. G. Shivakumar and P. Georgiou, "Transfer learning from adult to children for speech recognition: Evaluation, analysis and recommendations," arXiv preprint arXiv:1805.03322, 2018.

[3] R. Tong, L. Wang, and B. Ma, "Transfer learning for children's speech recognition," in 2017 International Conference on Asian Language Processing (IALP), Dec 2017, pp. 36–39.

[4] K. Okabe, K. Takafumi, and S. Koichi. "Attentive Statistics Pooling for Deep Speaker Embedding." In *Interspeech 2018*, 2252–56. ISCA, 2018.

Alberto Abad

Training Network on Automatic Processing of PAthological Speech (TAPAS)

phdopendays.tecnico.ulisboa.pt