Articulatory normalization via imitation strategy in phone classification task

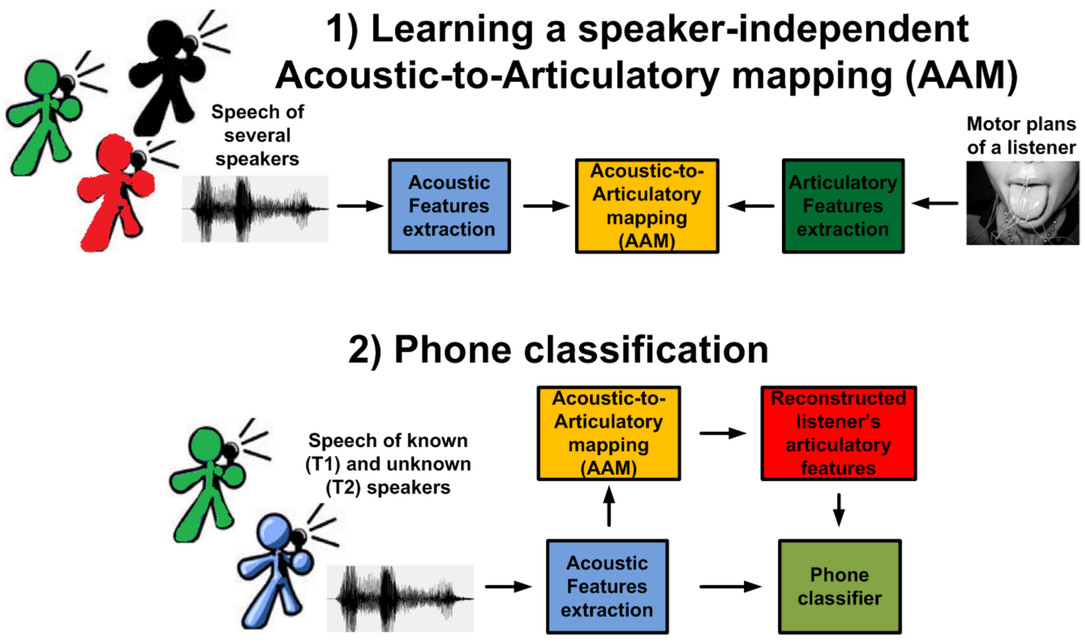
PhD Student Claudia Canevari
Leonardo Badino, Luciano Fadiga, Giorgio Metta
IIT – Istituto Italiano di Tecnologia
RBCS – Robotics, Brain and Cognitive Sciences Department

Introduction

Automatic Speech Recognition (ASR) systems classically apply normalization strategies that minimize inter-speaker variability by explicitly removing speakers' peculiarities ([1], [5]) or adapting different speakers to a reference model ([7]) or creating a compact adaptable and robust models ([2]) without exploiting speakers' variations in learning and recognition processes.

Approach

We propose a speaker normalization strategy that uses measured articulatory information and test it in a phone classification task ([4]).



Questions

1) Does the Articulatory normalization produce a more accurate phone classification w.r.t.

- no normalization strategy ?
- a corresponding acoustic normalization strategy?

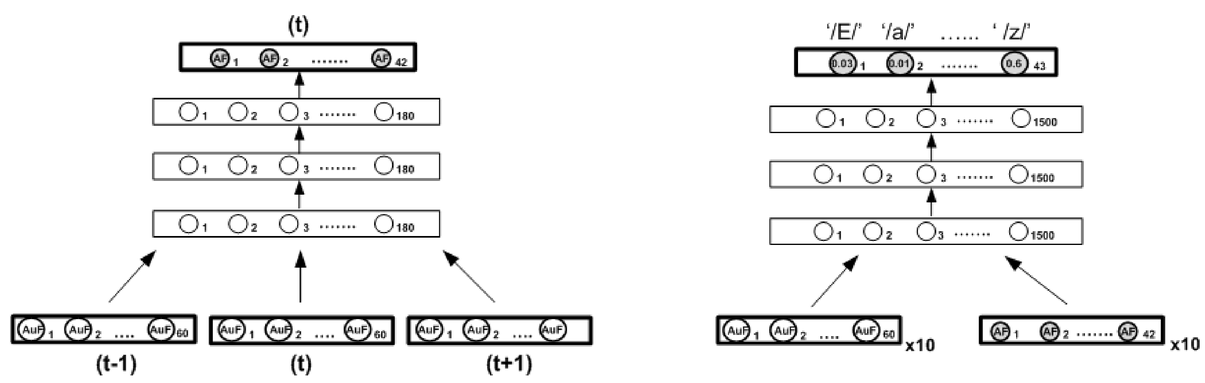
2) How can articulatory similarities affect the phone discrimination?

Material and Methods

Dataset: Italian corpus of simultaneous recording of speech and the trajectories of 7 articulators (UI, LI, LL, UL, TT, TB, TD) tracked by electromagnetic articulography. It includes 3120 single word utterances pronounced by 5 female subjects ([6]).

Acoustic-to-Articulatory Mapping* — 4-layer DNN ([3])

Phone classifier

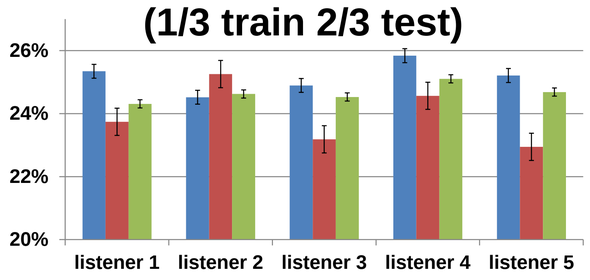


**An identical DNN is also used during acoustic normalization to perform Acoustic-to-Acoustic Mapping.*

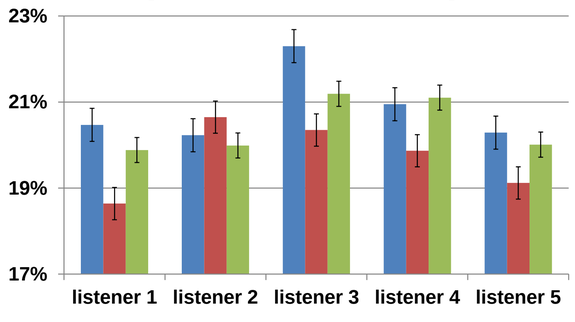
Results

1) Phone Classification Error

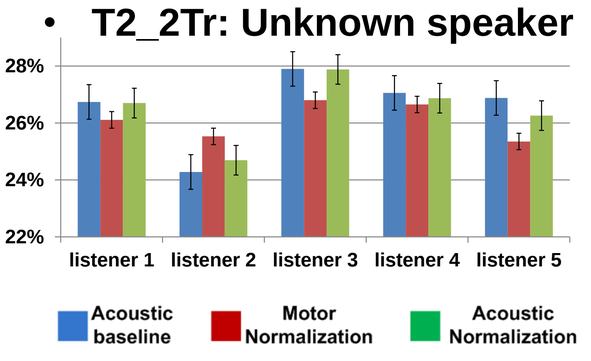
- **T1_1Tr: Known speaker (1/3 train 2/3 test)**



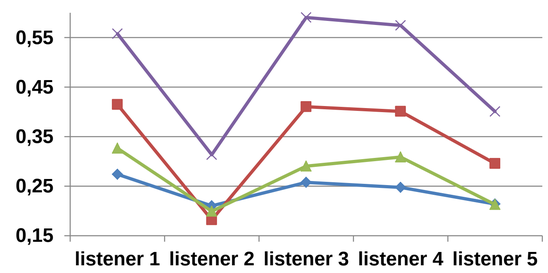
- **T1_2Tr: Known speaker (2/3 train 1/3 test)**

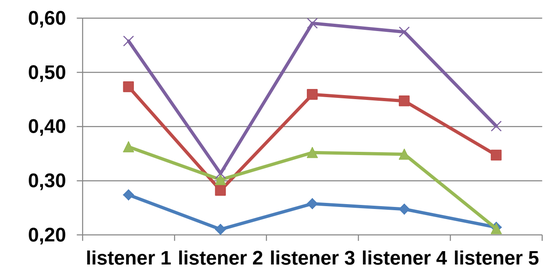


- **T2_2Tr: Unknown speaker**



Pearson Correlations





2) Articulatory similarities

	Correlation (pval)	
PER reduction/	**T1_1Tr**	**T1_2Tr**
Pearson corr	**48%** (0.032)	42.5% (0.06)
Pearson corr	**54.5%** (0.013)	41.9% (0.07)
Pearson corr	28% (0.23)	4% (0.8)

References

[1] Anastasakos, T., Kubale, F., Makhoul, J., Schwartz, R., "Adaptation to new microphones using tied-mixture normalization, in Proc. of IEEE International Conference on Acoustics", Speech and Signal Processing, Adelaide, SA, 1994, pp. 433-436.
[2] Anastasakos, T., McDonough, J., Schwartz, R., Makhoul, J., "A Compact Model for Speaker-Adaptive Training", in Proc. of Spoken Language Processing, Philadelphia, PA, 1996, pp. 1137-1140.
[3] Badino, L., Canevari, C., Fadiga, L., Metta, G., "Deep-Level Acoustic-to-Articulatory Mapping for DBN-HMM Based Phone Recognition",, in Proc. of IEEE spoken language technology workshop, 2012, Miami.
[4] Canevari, C., Badino, L., D'Ausilio, A., Fadiga, L., Metta, G., "Modeling speech imitation and ecological learning of auditory-motor maps", Frontiers in psychology, June 2013, 4.
[5] Eide, H., Gish, H., "A parametric approach to vocal tract length normalization", in Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing, Atlanta, GA, 1996, pp. 346-349.
[6] Grimaldi, M., Gili Fivela, B., Sigona, F., Tavella, M., Fitzpatrick, P., Craighero, L., et al., 'New tecnologies for simultaneous acquisition of speech articulatory data : 3D articulograph, ultrasound, and electroglottograph", in Proc. of Language Teching, 2008, Rome, Italy.
[7] Huang, X., Acero, A., Hon, H., W., "Spoken Language Processing", Upper Sandle River, New Jersey, NJ: Prentice-Hall.