Assessing phonological learning in COSMO, a Bayesian model of speech communication

Marie-Lou Barnaud, Jean-Luc Schwartz, Julien Diard, Pierre Bessière

Assessing phonological learning in COSMO, a Bayesian model of speech communication

Marie-Lou Barnaud
and Jean-Luc Schwartz
Univ. Grenoble Alpes, CNRS, Gipsa-lab
UMR 5216, F-38000 Grenoble, France
marie-lou.barnaud@gipsa-lab.grenoble-inp.fr

Julien Diard
Univ. Grenoble Alpes,
CNRS, LPNC UMR 5105,
F-38000 Grenoble, France

Pierre Bessière
CNRS, SORBONNE Universités
UPMC, ISIR, Paris, France

I. Introduction

According to speech development data, prelexical infants are able to discriminate and categorize syllables very early, almost from birth [1]. By contrast, categorization of phonemic units, especially consonants, arises probably not before nine months, in relation with speech motor development [2].

Learning processes involved in syllabic and phonemic categorizations are still unclear. This is why we propose a phonological version of our sensory-motor Bayesian model COSMO [3] to investigate this issue. In agreement with data on speech development [4], we consider that learning involves a sensory phase followed by a sensory-motor phase. In this paper, we show preliminary results obtained with this framework. We study the categorization of syllables, vowels and consonants after the first phase only, and after both phases. We compare performance at these two learning stages, especially with respect to the emergence of vowel and consonant invariant [5].

II. The phonological COSMO model

A. Model description

COSMO is a generic sensory-motor model of speech communication based on a Bayesian framework [6]. It deals with five variables: a sensory representation S, a motor representation M, two linguistic categorical objects O_L (L for Listener) and O_S (S for Speaker), respectively linked to S and M, and a Boolean switch C ensuring the coherence between both objects. The joint distribution $P(C\ O_L\ S\ M\ O_S)$ can be composed of elementary probability distributions: a prior on objects $P(O_S)$ and $P(O_L)$, a motor repertoire $P(M \mid O_S)$, an internal forward model $P(S \mid M)$, an auditory repertoire $P(S \mid O_L)$ and a validation system $P(C \mid O_S\ O_L)$ [1].

In the phonological model called COSMO-SylPhon, we consider Consonant-Vowel (CV) syllables, and we assume that the model includes both a syllabic decoding process with objects O^{Syl}, and a phonemic decoding process with objects O^C and O^O respectively for the consonant (represented by a closed configuration C) and the vowel (represented by an open configuration O). Syllables are characterized by a motor repertoire $P(M^C\ M^O \mid O_S^{Syl})$ and an auditory repertoire $P(S^C\ S^O \mid O_L^{Syl})$. Motor and sensory spaces M and S are themselves related by forward internal models $P(S^C \mid M^C)$ and $P(S^O \mid M^O)$. Furthermore, since classically the consonant motor gesture depends on the vowel motor gesture (which is called "coarticulation"), we also add a motor dependency $P(M^C \mid M^O)$. Finally, we compare this syllabic model to consonant and vowel models respectively characterized by motor repertoires $P(M^C \mid O_S^C)$ and $P(M^O \mid O_S^O)$ and auditory repertoires $P(S^C \mid O_L^C)$ and $P(S^O \mid O_L^O)$. A schema of the dependency structure of the syllabic and the phonemic models is displayed in Fig. 1.

Motor and auditory repertoires are implemented as Gaussian Mixture distributions in which each Gaussian distribution corresponds to one object. At the beginning, each Gaussian distribution approximates a uniform distribution and we expect them to gather around phonological units in the course of the learning process.

B. Learning

These distributions are learned in two phases, sensory then sensory-motor, in an iterative unsupervised process. At each step, a Master Agent produces a syllable resulting in an unlabelled sound s composed of the couple (s^c, s^o). During sensory learning, the Learning Agent learns its syllabic auditory repertoires: it selects the best Gaussian component o^{syl} corresponding to the sound s thanks to $P(O_L^{Syl} \mid [S = s])$ and then updates its distribution $P(S \mid O_L^{Syl})$ with the couple (s, o^{syl}). In the same way, it updates its consonant auditory repertoire thanks to a sound s^c and a consonant o^c and its vowel auditory repertoire thanks to a sound s^o and a vowel o^o.

During sensory-motor learning, the Learning Agent learns its internal model and its motor repertoires. As the Master Agent only provides a sound s, the Learning Agent first infers a motor gesture m, composed of the couple (m^c, m^o), thanks to

[1] The Boolean switch and the validation system are not taken into account in this study.

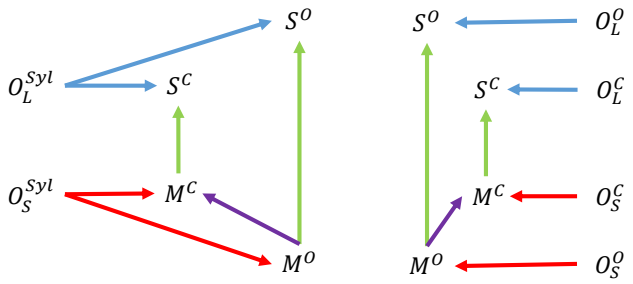

Fig. 1. The schematic phonological COSMO-SylPhon models. Left: the syllabic model. Right: the phonemic model. Each color line corresponds to a piece of knowledge of the model: motor repertoires in red, internal models in green, auditory repertoires in blue and motor dependency in purple.
O_L & O_S: categorical objet, S: sensory representation, M: motor representation; Syl: syllables, O: open (vowel), C: close (consonant)

$P(M \mid [S = s])$. Then, it updates its internal model $P(S|M)$ with the couple (m, s) and, in parallel, it uses its motor gesture m to select the best Gaussian component o^{syl} thanks to $P(O_S^{Syl} \mid [M = m])$. Finally, it updates its syllabic motor repertoire $P(M \mid O_S^{Syl})$ with the couple (m, o). In the same way, it updates its consonant motor repertoire thanks to a motor gesture m^c and a consonant o^c and its vowel motor repertoire thans to a motor gesture m^o and a vowel o^o.

C. Perception and production processes

The Learning Agent perception process involves a decoding distribution $P(O|S)$ from stimulus sound S to object O. The nature of the decoding depends on the choice of object O. For auditory decoding, based on sensory information, the selected distribution is $P(O_L|S)$. For motor decoding, based on motor information, the distribution is $P(O_S|S)$. Sensory learning enables to develop auditory decoding while sensory-motor learning enables to develop motor decoding.

In the following, two production processes are also tested: an imitation process $P(M|[S = s])$ to repeat sound s and a communication process $P(M|[O_S = o])$ to produce object o.

III. Results

After learning, the model is evaluated in two ways. Firstly, we assess learning with a sound imitation task during which the Learning Agent has to repeat perceived sounds s produced by the Master Agent. This results in sounds s' that we compare to s to compute the average Kullback-Leibler divergence. As we evaluate both syllabic and phonemic performance, s and s' can be respectively S^{Syl}, S^C and S^O. We observe in each case that divergence is close to zero.

Secondly, we assess the categorization ability for which the Learning Agent perceives and tries to repeat objects o produced by the Master Agent as sounds s. This repetition results in sounds categorized as objects o' by the Master. We finally compute the confusion matrix between objects o and o'. In our model, this task combines auditory or motor decoding with the communication process. Objects o and o' can be syllables O^{Syl} or phonemes O^C and O^O. With auditory decoding, we observe in the confusion matrix that syllables and vowels can be correctly decoded but not consonants, due to consonant overlaps. However, with motor decoding, consonants are better decoded supporting the assumption that phonemic categorization could require sensory-motor learning.

IV. Conclusion

In this paper, we propose a phonological version of the COSMO model based on a sensory-motor framework. We focus on categorization ability and we present a learning process in agreement with the developmental data.

Acknowledgment

Research supported by a grant from the European Research Council (FP7/2007-2013 Grant Agreement no. 339152, "Speech Unit(e)s").

References

[1] P. Hallé and A. Cristia, "Global and detailed speech representations in early language acquisition," in *Speech production and perception: Planning and dynamics*. Peter Lang, 2012. [Online]. Available: http://pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:1381615:4

[2] M. Dole, H. Loevenbruck, O. Pascalis, J.-L. Schwartz, and A. Vilain, "Phoneme categorization depends on production abilities during the first year of life," *ICIS 2016 - The International Congress on Infant Studies*, 2016.

[3] C. Moulin-Frier, J. Diard, J.-L. Schwartz, and P. Bessière, "COSMO ("Communicating about Objects using Sensory-Motor Operations"): A Bayesian modeling framework for studying speech communication and the emergence of phonological systems," *Journal of Phonetics*, vol. 53, pp. 5–41, 2015.

[4] P. K. Kuhl, "Early language acquisition: cracking the speech code," *Nature Reviews Neuroscience*, vol. 5, no. 11, pp. 831–843, Nov 2004. [Online]. Available: http://dx.doi.org/10.1038/nrn1533

[5] R. Laurent, M.-L. Barnaud, J.-L. Schwartz, P. Bessière, and J. Diard, "A computational assessment of the role of motor information in speech perception," *Psychological Review*, in press.

[6] P. Bessière, E. Mazer, J. M. Ahuactzin, and K. Mekhnacha, *Bayesian Programming*. Boca Raton, Florida: CRC Press, 2013.