Downloaded from UvA-DARE, the institutional repository of the University of Amsterdam (UvA)
http://hdl.handle.net/11245/2.53355

File ID	uvapub:53355
Filename	ICPhS_normalization_final.pdf
Version	unknown

SOURCE (OR PART OF THE FOLLOWING SOURCE):

Type	conference contribution
Title	Modelling vowel normalization and sound perception as sequential processes
Author(s)	P. Escudero, R.A. Hoffmann Bion
Faculty	FGw: Amsterdam Center for Language and Communication (ACLC)
Year	2007

FULL BIBLIOGRAPHIC DETAILS:
http://hdl.handle.net/11245/1.275788

Copyright

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content licence (like Creative Commons).

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
(pagedate: 2014-11-17)

MODELING VOWEL NORMALIZATION AND SOUND PERCEPTION AS SEQUENTIAL PROCESSES

Paola Escudero and Ricardo Augusto Hoffmann Bion
University of Amsterdam
paola.escudero@uva.nl, ricardobion@gmail.com

ABSTRACT

This study constitutes the first attempt at combining vowel normalization procedures with the linguistic perception framework of Stochastic Optimality Theory [1] and the Gradual Learning Algorithm [2]. Virtual learners possessing different normalization procedures, and a control learner with no normalization, were trained to perceive Brazilian Portuguese and American English vowels. Our results show that learners equipped with normalization algorithms outperformed the control learners, obtaining accuracy scores up to 33% higher. Thus, this model in which normalization and sound perception are implemented as two sequential processes seems to be able to explain sound categorization adequately. That is, it improves the performance of a perception grammar when the training and testing sets have speakers with different ages and gender.

1. INTRODUCTION

Listeners seem to build their perceptual systems on the basis of the statistical distribution of the acoustic stimuli they hear [1, 3]. However, accurate vowel perception cannot emerge exclusively from these distributions because acoustic cues for vowel identity vary greatly among the productions of different speakers, due to both anatomical/physiological and sociolinguistic reasons [4]. To account for the fact that listeners manage to recognize words spoken by different speakers despite this variation, several vowel normalization procedures have been proposed [for reviews, see 5, 6].

Despite this variability in the speech signal, several studies have simulated the learning of vowel perception from production data without incorporating vowel normalization [7, 8, 9]. Instead, data from a single speaker or hypothesized distributions were used for both the training and testing of models. This, of course, raises questions as to the validity of the model in a natural environment characterized by multiple speakers.

The present study thus aims at bridging this gap by proposing a sequential model of vowel perception in which normalization precedes linguistic processing. We also model linguistic perception within the framework of Stochastic Optimality Theory (Stochastic OT) [1] and the Gradual Learning Algorithm (GLA) [2], following [7, 8], while vowel normalization is instantiated by different procedures available in the literature.

2. VOWEL PERCEPTION AND STOCHASTIC OPTIMALITY THEORY.

The main assumption of a perception grammar modeled within Stochastic OT is that phonetic categories derive from auditory inputs after the resolution of conflicts among *cue constraints* [1, 8]. These constraints are continuously ranked by their order of importance, and the optimal output of the grammar is the candidate that causes the least serious violations of constraints. When there is a mismatch between the intended input and the output of the grammar, *lexicon-driven* perceptual learning occurs in the form of constraint re-rankings so as to avoid future errors [1, 8].

The acquisition of vowel perception has been modeled by [7] on the basis of two auditory cues, namely F1 and F2. In this study, vowel categorization consisted of mapping auditory inputs (e.g. F1 = 250, F2 = 1900) to vowel categories (e.g. /i/ or /e/), through the resolution of negatively formulated cue constraints such as "an F1 of 360 Hz is not /i/". A tableau with only two candidates and the learning process is illustrated in Tableau 1.

Tableau 1. Lexicon-driven perceptual learning

F1= 250Hz F2=1900Hz [i]	F1=250 not /i/	F1=250 not /e/	F2=1900 not /i/	F2=1900 not /e/
√ /i/	*!⇒		*⇒	
☞ /e/		⇐*		⇐ *

In the tableau above, the auditory input [F1=250Hz, F2=1900Hz], intended as the vowel

/i/, was incorrectly identified as the vowel /e/, because, at this time of the evaluation, a learner had the constraint "F1=250Hz not /i/" as the highest ranked. Here, the lexicon can act as a supervisor, which tells the learner she committed a mistake in vowel categorization because the speaker intended a word containing the vowel /i/. As a result, constrains can be re-ranked, as shown by the arrows, so that this sort of "error" is less likely to occur in the future.

2.1. Integrating normalization into the model

Here, we follow the learning procedure outlined above with the addition of a pre-linguistic normalization module. We do not discuss how the normalization algorithms are learned, but rather we implement different normalization procedures available in the literature [6], and assume they have been mastered before lexicon-driven learning occurs. The early development of vowel normalization procedures is supported by experimental studies which indicate that infants can recognize similarities among vowels produced by different speakers, i.e. normalize them, before language-specific vowel categories emerge [10]. Additionally, our modular view of normalization is practical for simulation purposes, as speaker-dependent variation is eliminated before categories are acquired or vowels are categorized.

2.1.1. The normalization module

Table 1 shows the four virtual learners we simulated in this study, three equipped with normalization procedures (LOBANOV [11], NEAREY1 [12], and GERSTMAN [13]), and a control learner, with no normalization procedure.

Table 1. Virtual learners.

Listener	algorithm
1. HZ	Control with no normalization algorithm
2.LOBANOV	Lobanov's z-score transformation
3.NEAREY1	Nearey's single logmean procedure
4.GERSTMAN	Gerstman range normalization:

We have chosen not to include any vowel-intrinsic normalization procedure in our simulations because previous studies [6] have shown that these perform much worse than the extrinsic procedures here considered.

In the proposed normalization module, the virtual learners normalized the F1 and F2 values present in the training and testing sets according to the normalization algorithm with which they were equipped.

Importantly, for vowel-extrinsic algorithms to work, they require information about the vowel space of the speaker whose data is to be normalized. That is, LOBANOV requires information about the mean and standard deviation (sd) of a given formant, NEAREY1 requires the mean across log-transformed formant values, and GERSTMAN requires information about the maximum and minimum values of the formants produced by a speaker. This information should be derived from a large set of formant values produced by the speaker to be normalized. For the LOBANOV algorithm, for instance, if F1 values of a given speaker are to be normalized, the mean F1 and the sd of this mean across several vowel tokens of the same speaker need to be computed in advance.

We chose to pre-compute these pieces of information from our datasets, and to give these pieces of information to our virtual learners from their very first contact to a new speaker. This choice is partially justified by the fact that, when confronted with a new speaker, listeners are likely to have heard speakers with similar voices and can probably use this information for vowel normalization. Given that we simulate lexicon-driven learning of vowel perception, which follows the acquisition of vowel categories, we can safely assume that, at this point in their development, listeners have already been in contact with hundreds of different speakers.

Once these pieces of information were computed, the algorithms work in the following way: LOBANOV [11] normalizes any incoming formant value by subtracting from it the mean formant value it computed and by then dividing this result by the standard deviation of the mean. NEAREY1 [12] normalizes an incoming formant value by subtracting from its logarithm the average of the logarithms of the same formant of the same speaker. GERSTMAN [13] normalizes an incoming formant value of a certain speaker by subtracting from it the minimum value of the same formant of the same speaker, and by then dividing the result by the difference between the maximum and minimum values of the same formant of the same speaker.

Thus, our *normalization module* simply normalizes any F1 and F2 value by transforming them with algorithms available in the literature. Crucially, all of these algorithms use only the F1 and F2 values of each token produced by each speaker without having access to any linguistic information about them, e.g. their phonological labels. In our modelling, after this first normalization process occurred, the acoustically-normalized values were fed into the *linguistic perception module*.

2.1.2. The linguistic perception module

This module follows the procedure of [7]. The input of the grammar, and its constraints and tableaus are established according to the 14 F1 and 10 F2 values defined in the *normalization module*. Thus, our perceptual grammar consists of 168 constraints in the Brazilian Portuguese (BP) simulations and 264 in the American English (AE) simulations, i.e. (10 F1 values + 10 F2 values) × (7 BP or 11 AE vowels), and 140 tableaux, containing all possible combinations of F1 and F2 values in the environment (14 F1 × 10 F2 values).

In order to train our virtual learners, we fed them with 400,000 F1 and F2 combinations drawn from the training set. If the learner was equipped with a normalization algorithm, these values had been previously normalized with its corresponding normalization algorithm within the *normalization module.* In addition to the formant values, the label of the vowel intended by the speaker was also fed to the grammar, so that constraints could be re-ranked in the event of inaccurate categorization. At the onset of learning, all constraints had the same *ranking value* of 100. *Plasticity*, which is the speed by which constraint rankings can change, decreased from 1.0 to 0.001 throughout learning, and the *evaluation noise* of the grammar was kept constant at 2.0. At the end of the training phase, the virtual listeners' perception grammar had a ranking of constraints that reproduced the relation between auditory inputs (i.e. F1 and F2 combination), and vowel categories (e.g. /i/, /a/) found in the environment (training set).

During the testing phase, our virtual listeners categorized 100.000 auditory inputs drawn from the testing set. As in the learning phrase, if listeners were equipped with normalization algorithms, these auditory inputs were processed by the *normalization module* before being fed to the perceptual grammar. After normalization, the auditory inputs were categorized trough the perception grammar constraint rankings. Importantly, the vowels from the new speakers in the testing set were categorized on the basis of the raking of constraints acquired during the training phrase. Virtual learners' degree of accuracy in vowel perception was calculated by comparing their categorization of auditory inputs with the vowels originally intended by the speakers to give percent-correct identification.

3. RESULTS OF THE SIMULATIONS

3.1. Brazilian Portuguese Vowels.

Training and testing stimuli: Eight monolingual speakers (4 female, 4 male) recorded CVCV words containing the seven oral vowels of BP, resulting in a total of 20 productions of each vowel per speaker. Four speakers (2 female, 2 male) were included in the testing set, and four in the training set. The first two formants (F1 and F2) of the vowels were measured and used to train and test the model.

Results: The chance level of a given vowel being correctly identified is of 14% (100% / 7 vowels), thus, any improvement on this number indicates that the simulations were somehow efficient. Table 2 presents a summary of the performance of the 4 virtual learners.

Table 2. Percentage of correct labeling of the virtual learners

Listener	% accuracy
Chance level	14
1.HZ (control)	75
2.LOBANOV	90
3.NEAREY1	81
4.GERSTMAN	87

In Table 2, we see that virtual learner 1 (control learner) managed to reach 75% accuracy when labeling vowels from new speakers, despite not relying on any normalization algorithm. This is surprising given that this learner had been exposed to the vowels produced by only four speakers.

Learners 2, 3, and 4, who were equipped with normalization procedures, however, outperformed the control. Learner 5 (Lobanov algorithm) performed the best, reaching 90% correct. Thus, the inclusion of pre-linguistic normalization procedures improved the performance of the perceptual model. Specifically, it can result in up

to a 25% an increase in learners' correct vowel identification.

We hypothesize that this difference between the control learner and the learners with normalization algorithms would increase if the language to be learned contained a greater number of vowels, and more overlap in the vowel plane. Thus, we conducted a new simulation in which listeners had to learn the 11 AE vowels, rather than the 7 BP.

3.2. American English Vowels.

Training and testing stimuli: Eight monolingual speakers (4 female, 4 male) recorded CVC words containing the eleven vowel monophthongs of AE, resulting in a total of 15 productions of each vowel by each speaker. The rest of the procedure follows the description in 3.1.

Results: Table 3 shows that, just like in the previous experiment, listeners with normalization algorithms performed the best.

Table 3. Percentage of correct labeling of the virtual learners

Listener	% accurate
Chance level	09
1.HZ (control)	49
2.LOBANOV	72
3.NEAREY1	65
4.GERSTMAN	57

We see that the learner equipped with the Lobanov algorithm again had the best accuracy scores, 72% of correct labeling. This is way above chance level of 9% correct, and it is 33% higher than the correct responses of our control. This supports the hypothesis that normalization algorithms become more important with an increase in the number of vowels to be learned. The 72% of accuracy of our best learner might appear to be a low value, but it is instead promising, as learners have only been exposed to 60 tokens of each vowel during their training phase. Also, only F1 and F2 values were given to the model during training and testing, despite the role that, for instance, vowel duration and spectral changes plays in the perception of AE vowels.

4. DISCUSSION AND CONCLUSIONS

As shown in this paper, a model in which normalization and sound perception are implemented as two sequential processes adequately accounts for sound categorization. That is, normalizing formant values before feeding them to a linguistic perception grammar greatly increases the performance of this grammar. In both our simulations, virtual listeners equipped with normalization algorithms outperformed the control learner not endowed with normalization algorithms. These findings are particularly encouraging, as only two acoustic cues were given to the learners, and our speakers on the training and testing phases had different ages and gender. As found in previous studies [6], the Lobanov algorithm appears as the best algorithm to reduce between-speakers differences in vowel production.

5. REFERENCES

[1] Boersma, P. (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.

[2] Boersma, P. & Hayes. B. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32: 45–86.

[3] Kuhl, P. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience, 5*, 831–843.

[4] Peterson, G., & Barney, H. (1952), Control methods used in a study of the vowels, *J.Acoust.Soc.Am. 24,* 175-184

[5] Rosner, B., & Pickering, J. (1994). *Vowel Perception and Production*. Oxford: Oxford University Press.

[6] Adank, P., Smits, R., & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research, *J.Acoust.Soc.Am. 116*(5), 3099–3107.

[7] Boersma, P. & Escudero, P. (2004): Learning to perceive a smaller L2 vowel inventory: An Optimality Theory account. *ROA* 684

[8] Escudero P. & Boersma, P. (2004): Bridging the gap between L2 speech perception research and phonological theory. *SSLA,* 26, 4: 551-585.

[9] Boersma, P., Escudero, P. & Hayes, P (2003) Learning abstract phonological from auditory phonetic categories. *Proc.15th ICPhS*, Barcelona. pp. 1013-1016.

[10] Jusczyk, A. (1997). *The discovery of spoken language*. Cambridge: The MIT Press.

[11] Lobanov, B. (1971). Classification of Russian vowels spoken by different speakers. *J.Acoust.Soc.Am., 49*, 606–608.

[12] Nearey, T. (1978). *Phonetic Feature Systems for Vowels*. Indiana. Indiana University Linguistics Club.

[13] Gerstman, L. (1968). Classification of self-normalized vowels. *IEE , AU-16*, 78-80.

ACKNOWLEDGEMENTS

The authors thank Dr. Paul Boersma, from the University of Amsterdam, and Dr. Geoffrey Morrison, from Boston University, for their comments on earlier versions of this paper. The usual disclaimers apply.