Interspeech 2018
2-6 September 2018, Hyderabad

Category Similarity in Multilingual Pronunciation Training

Jacques Koreman

Norwegian University of Science and Technology, Trondheim, Norway
jacques.koreman@ntnu.no

Abstract

Learners with different native languages (L1) meet different challenges when they learn a foreign language (L2). The Speech Learning Model and the Perceptual Assimilation Model PAM-L2 have led to important insights about these challenges. Among other things, they have shown that the learnability of L2 sounds depends on their similarity to sounds in the L1: L2 sounds are more likely to lead to the formation of new phonetic categories if they differ strongly from L1 categories than if they are similar. The similarity of sounds is hard to quantify objectively, especially if the aim is to do this for many L1-L2 pairs. This limits the models' practical applicability.

The multilingual pronunciation training platform CALST offers exercises for all new L2 sounds. Two implementations of category (dis)similarity are proposed to identify new sounds, one at the level of functional similarity maintaining all L2 phonemic contrasts, the other based on a more fine-grained, multilingual similarity measure, where L2 sounds are considered new if they can contrast phonemically with the most similar L1 sound in *any one* language. This level of granularity reflects phonetically salient differences between sounds which, when perceived and produced adequately, suffice for high intelligibility and comprehensibility in L2.

Index Terms: multilingual pronunciation training, phonetic similarity, functional similarity, multilingual similarity

1. Acquiring L2 sounds

When learning a new language, perceiving and producing the sounds of the foreign or second language (L2) is important for successful and effective communication. This does not necessarily mean that perception and production must be identical to those of a native speaker, but all L2 sound contrasts should be perceived correctly by the learner, and the learner's own productions of those contrasts should be comprehensible to a native listener.

The multilingual pronunciation training platform CALST (Computer-Assisted Listening and Speaking Tutor) which is discussed in this article, is designed as a learning tool for both second and foreign language learning situations, both as a complement to classroom teaching and for individual use [1]. It is aimed at adult learners at the beginner level, so that we can assume that the sound system of the learners' native language (L1) is fully developed, while their level of L2 learning is similar to that of naïve listeners, i.e. they are functional monolinguals with little or no previous knowledge of the L2. The learners are thus assumed to be in a stage where most perceptual learning takes place [2, p. 20]. In the two following sections, the two most influential L2 learning models, PAM-L2 and the SLM, will be discussed, together with some of the limitations for their application in a multilingual learning platform.

1.1. PAM-L2

The Perceptual Assimilation Model PAM-L2 for second language acquisition [2] distinguishes several scenarios for the perception of L2 speech sounds depending on their perceived articulatory similarity to speech sounds in the learner's L1. Sounds which are phonetically and phonologically similar to L1 sounds are assimilated to the L1 category (Categorized). Other sounds may be perceived as speech sounds which are not similar to any L1 sound (Uncategorized) or even as non-speech sounds (Non-Assimilated). Depending on the categorization of the two sounds in cross-language contrasts, six assimilation types are defined. The most difficult contrast occurs when two L2 sounds are equally good exemplars of the same L1 category; such a contrast will be hard to learn (e.g. /ɓ-b/ in L2, if L1 only has /b/). At the other end of the discriminability scale, if the two non-native speech sounds are acceptable exemplars of two different L1 categories, very good or excellent discrimination is expected. This has been attested for the perception of the English [w-ɫ] contrast by German learners, who assimilated these sounds to the L1 phones /v/ (usually realized as an approximant [ʋ] in German) and /l/, respectively [3].

Being a perception model, PAM-L2 makes important claims about the perceptual categorization of L2 speech sounds in terms of the closest L1 sounds. If we want to use the model to predict communication problems in a foreign/second language, however, the picture is complicated considerably. Since communication between a learner and a native speaker is a two-way process, we not only have to consider the perception of L2 sounds by the learner, but also the way the learner's productions of the assimilation categories are perceived by a native speaker. In the example of the [w-ɫ] contrast above, a German realization of syllable-final [ɫ] as [l], using a clear instead of a dark /l/, probably will not cause any problems for an English listener: although the selection of this allophone in syllable-final positions is inappropriate and betrays the speaker's foreign accent, an English listener will perceive it as a realization of the phoneme /l/ without any problems. This is not so obvious for the use of the sound [v] (or [ʋ]) instead of [w], however. Even though the realization of the German sound differs from its closest English counterpart [v], the German sound will probably be identified as /v/ by native listeners and not as /w/. (The situation is less clear for the more usual German realization as [ʋ], which is labiodental like [v], but an approximant like [w].) So even though the [w-ɫ] contrast is not difficult to perceive for German learners, it can cause communication problems.

Extending PAM-L2 to two-way communication in this way, there are many Two Category assimilations which are symmetrical, i.e. where the two perceived L1 categories map back onto the original or intended contrast in the L2 when they are produced by the learner. An example would be the German and English speech sounds /i:/ and /u:/, even though they are

10.21437/Interspeech.2018-1938

pronounced somewhat differently in the two languages. These are assumed to be unproblematic. We predict that comprehensibility will be lower when the sounds in a contrast are poor category exemplars: The contrast between English /l/ and /r/ will generally not lead to confusions in a German L2 learner's perception, but the very different phonetic realizations of the /r/ sound in German and English may lead to lower comprehensibility across the two languages. We are therefore suggesting that Two Category assimilations which lead to very good or even excellent discrimination can nevertheless cause communication problems in L2. Comprehensibility of German learners by native speakers of English is expected to decrease going from [i:-u:] with two fairly good German exemplars of the corresponding English vowel categories to [l-r] with one poor exemplar to [w-ł] with one L2 sound realization categorized as another phoneme in English, which we call an asymmetric Two-Category assimilation. Note that this assimilation label is used to signify L2 communication here, as an extension of their original use which is limited to L2 perception.

Taking into account native speakers' perception of the phone realizations by L2 learners puts a stronger focus on the learner's articulatory capabilities. Uncategorized L2 sounds and especially Non-Assimilated sounds may pose big problems for a learner: Even though such sounds may be easy to distinguish perceptually from Categorized L2 sounds, they are probably difficult to produce. Remember that PAM-L2 assumes that a learner perceives distal articulatory events. Especially for sounds where these distal events are outside but similar to the learner's L1 system, as for Uncategorized speech sounds, or even outside his/her language experience, as for Unassimilated sounds, it is reasonable to assume that much more effort is required to perceive the distal articulations and use this knowledge to produce these unfamiliar sounds. Untypical L2 productions of these sounds are likely to reduce their comprehensibility to native listeners, so that Uncategorized-Categorized (UC), Non-Assimilated (NA) and Uncategorized-Uncategorized (UU) assimilations may cause bigger problems in communication than they do in L2 perception (cf. [3]). This remains to be evaluated in further investigations.

CALST offers discrimination (ABX) and identification exercises using L2 minimal pairs to direct the beginning learner's attention to the differences between similar speech sounds, which is especially important for Single Category assimilations. Because the minimal pair exercises contrast unfamiliar sounds with *all* similar sounds in the L2, the learner will also have to deal with asymmetrical Two Category assimilations: Even if a German learner of English (above) has no problem at all distinguishing English /w/ from /ɫ/, the first sound is contrasted with /v/ in another exercise, where it represents a Single Category assimilation. By offering listening and pronunciation exercises for L2 contrasts between all similar L2 sounds, for instance all contrasts in which the two sounds in a pair differ in only one dimension in the IPA consonant or vowel charts, CALST aims to guarantee intelligibility and comprehensibility of L2 learners' productions in their communication with native speakers.

1.2. SLM

According to the Speech Learning Model (SLM), learners "perceptually relate positional allophones in the L2 to the closest positionally defined allophone (or 'sound') in the L1" [4], called diaphones in [5, cited in 4]. The SLM relates L2 perception to the learner's productions. Like PAM-L2, the SLM assumes that unfamiliar L2 sounds are more easily learned if they are phonetically different from any sound in the L1, while strong similarity of an L2 sound to a speech sound in the native language will lead to phonetic equivalence. The SLM posits that, given enough input, learners should in principle be able to learn any new sound, although there is a negative correlation with the Age of Onset of Learning (AOL). It should be noted that this does not imply that learners learn to speak the L2 without a foreign accent, even if the L2 is acquired at a young age: The L2 exists in the same phonetic space as the L1, and the SLM provides evidence that two languages influence each other in a learner. Phonetically similar phones are merged into equivalence categories with properties that are intermediate between the L2 and the L1 (e.g. VOT for French and English voiceless plosives). When L2 and L1 phones are merged, this also causes interference by which L1 sounds move towards the realization in L2, causing them to deviate from the L1 norm; this is called "reverse interference". Also, while new categories are easier to learn and are produced more precisely, they may not be accurate realizations of the L2 sounds which are identical to those of monolingual native speakers because dispersion pushes them away from existing L1 categories. This limits the learner's ability to achieve an accent-free pronunciation in the L2 (and in the L1), even though intelligibility and comprehensibility can be very high.

Like PAM-L2, SLM uses the concept of similarity to explain when a new category is formed by the L2 learner. But testing hypotheses with regard to the differences between L2 and L1 speech sounds is difficult because of "the lack of an objective means for gauging degree of perceived cross-language phonetic distance" [4, p. 264]. Similarity can only be measured through perception experiments in which speech sounds in L2 and L1 are compared directly. This is sometimes done for a selected set of speech sounds, in which listeners hear an L2 phone and classify it in terms of an L1 speech sound, rating the goodness of fit to the L1 sound [6]. A generalization of this approach to the comparison of the complete sound inventories of two languages would necessitate a large number of comparisons, since there may be several near neighbours to each L2 speech sound in the L1. This makes this approach unfeasible for the complete set of positional (allo)phones of two languages, let alone for all the languages in a multilingual pronunciation training platform. To avoid misunderstandings, let us stress that this argument is used to explain why this is not a viable approach to determining new categories in a multilingual pronunciation training platform like CALST; it does not in any way detract from the importance of such experiments for deepening our understanding of L2 learning. In addition to the scale problem, perceptual tests of L2 perception ignore the effect of L2 production on comprehensibility by native speakers, which would also need to be tested in identification and rating experiments.

2. Levels of similarity

In this section, we propose two levels of similarity, based on segment information that is available in L1-L2*map* [7], which is based on UPSID [8,9, see also 10]. This work is still in progress, and the aim of this article is to elicit feedback from L2 experts.

The two most influential models of L2 category learning make assumptions about the similarity of L2 phonetic (and phonological) categories to L1 speech sounds. A definition of similarity can in principle be obtained from perception tests as described in section 1.2, but this approach is not scalable to multilingual pronunciation training. Moreover, it is important to know which sounds lead to phonetic equivalence not just in L2 perception, but in two-way communication. That is, not only do L2 learners need to learn to perceive all contrasts in the L2, they also need to reproduce them in such a way that native speakers perceive them as the intended contrast in the L2. For L2 learning, this is important to achieve comprehensibility irrespective of the similarity of L2 and L1 categories. As a primary aim for beginning L2 learners, they must learn to hear and produce all L2 category contrasts, while it is acceptable if the realizations of the phonetic categories reveal a foreign accent. With increasing L2 experience, they will learn to hear and produce finer category distinctions.

To reflect progressive learning, we propose a practical, predictive approach (i.e. an approach which does not require previous knowledge obtained for instance from L1-L2 perception experiments) to determine the L2 speech sounds which must be learned as new phonetic categories in the L2. We shall propose two levels of similarity as a practical approach to L2 learning.

2.1. Functional similarity

A low level of granularity (cf. [11]) is defined by conflating all speech sound realizations into the sound categories defined in the IPA charts [12]. In accordance with the IPA consonant chart, for example, no distinction is made between dental, alveolar and postalveolar segments at this functional similarity level, except for fricatives and affricates. All language-specific phonetic realizations of the categories are thus replaced by their so-called base consonants, which correspond to the sounds shown in the IPA charts. Exercises are only triggered in CALST for base consonants in the L2 which do not occur in the learner's L1. It is important to note that speech sounds are only conflated into base sounds if there is no contrast between two variants of the base sound in the L2. For L2 Bulgarian, for example, palatalized and non-palatalized plosives are not conflated, because the contrast in the L2 must be maintained by the learner to achieve a minimum level of comprehensibility, which requires that all minimal pairs in the target language are distinguished by the learner.

Let us look at a few examples to evaluate the consequences of functional similarity. According to its UPSID description, French has a contrast between voiced and voiceless plosives at bilabial, dental and velar places of articulation. In Bulgarian, this plosive series is extended with palatalized versions. For French learners of Bulgarian, no exercises are selected for the non-palatalized plosives, which occur in both languages. But since the palatalized plosives are not conflated with the base consonants in order to maintain the contrast between non-palatalized and palatalized plosives in Bulgarian, French learners of Bulgarian do get exercises for the palatalized consonants. In the comparison of L1 Bulgarian with L2 French, the Bulgarian palatalized plosives are conflated with their non-palatalized counterparts. Since these base consonants also occur in French, no exercises are selected for Bulgarian learners of French. There is a silent assumption behind this procedure that Bulgarian learners of French will automatically pick the right L1 phoneme when they speak French, i.e. they will use the non-palatalized plosives. We return to this in the Section 3.

Let us stick to a comparison of plosives, and compare French with English. Initial (phonologically) voiceless stops in English are aspirated, while they are unaspirated in French. Initial voiced stops are also realized differently, namely with and without voicing in French and English, respectively. These differences are ignored when the speech sounds are conflated into base consonant series /p,b/, /t,d/ and /k,g/ which are found in the IPA consonant table, so that no exercises are selected for these sounds in either target language. In principle, the functional phonological contrast between voiced and voiceless consonants is maintained in both French and English as the L2, even though the series have very different realizations in the two languages. No exercises are therefore selected in CALST. But naïve English L2 learners of French may produce aspirated instead of unaspirated voiceless stops, revealing a strong foreign accent; and vice versa, French learners of English may produce pre-voiced stops, with the same effect. Simultaneously, the unaspirated, voiceless realization of /b/ in English and /p/ in French may lead to misperceptions. It is certainly known that aspiration is an important perceptual cue required for perception of a voiceless plosive in English in initial position, and voicing is likely to be important for the perception of a voiced plosive in French. We shall return to this in the next section.

Functional similarity thus does not necessarily require phonetic similarity. Best and Tyler [2] compare the French realization of /r/ as a voiceless uvular fricative [ʁ] with the English realization of /r/ as an alveolar approximant [ɹ]. When substituted in the L2, native listeners are aware of their functional similarities despite the phonetically different realizations of the rhotics: They have similar phonotactic and other (morpho)phonological properties, and they are also written identically in the two languages, which probably strengthens their perceived equivalence. Similarly, Spanish realizations of /r/ as [ɾ], as in 'pero' (E. *but*), or [r] as in 'perro' (E. *dog*), will be comprehensible as functional equivalents to a French or English native listener. But when French or English speakers learn to speak Spanish, the contrast between the two Spanish phonemes must of course be maintained. This can be implemented in L1-L2*map* by conflating all rhotics (fricative, approximant, tap or trill) when comparing two languages, as long as all contrasts in L2 (as in Spanish) are maintained.

We also use length as a segmental property, and we have included the IPA non-pulmonic and "other consonants" in the L1-L2*map* consonant table by adding rows for manner of articulation). Following Bohn [3, Table 1] implosives should probably be mapped to the same base consonants as their pulmonic counterparts, since they are considered as a Single Category assimilation with Spanish /b/; possibly, also ejectives should be mapped onto the same base consonants. Clicks, on the other hand, should not be mapped onto the corresponding non-pulmonic base consonants, since it is claimed they are non-assimilated [13].

Functional similarity ensures that all L2 contrasts are maintained by the L2 learner (if learning is successful). But as the above discussion shows, much of the burden of communication is placed on the native listener (cf. [14]), and much less on the learner – like in the example of the voiceless plosives produced by French learners of English and vice versa, or for the use of French and English /r/ across the two languages. Functional similarity therefore sets a lower bound on comprehensibility. For a smooth communication, more effort must be placed on the L2 learner. In the following section, we propose a much higher degree of granularity to operationalize cross-language similarity on the basis of available information in L1-L2*map* and UPSID.

2.2. Multilingual similarity

Greater phonetic similarity of a learner's productions of L2 categories to their realizations by native speakers reduces the communicative load on the native listener. A high level of granularity in the comparison of language pairs can be obtained by using a measure of similarity which reflects all *possible* phonemic oppositions in the languages of the world. Under this definition of similarity, any two speech sounds which can form a phonemic contrast in *any one* language are considered to be separate phonetic categories, with the underlying rationale that such categories reflect perceptually salient differences. Note that this multilingual similarity measure requires a definition of all sounds in any given language in phonetically accurate descriptions of the phoneme inventory of languages. The phonetic categories we use are based on the phonetic segments used to define the segment inventories in UPSID [8,9]. This principle is described by Ladefoged and Maddieson, who write that the phonetic description of sounds of the world's languages "needs to be rich enough to describe those segmental events which distinguish one language or accent from another and which are also sufficiently distinct to serve as potential conveyors of lexical contrasts for speakers of other languages" [15].

For instance, the different realizations of the /p-b/ contrast in French and English will be captured better when using "multilingual phones": The French voiceless plosives are still defined as unaspirated, while their English counterparts are now characterized as aspirated. The reason for this is that aspirated and unaspirated plosives are distinctive in for instance Thai, so that both speech sounds will be retained, even though [p] and [p^h] are not distinctive in either English or French. As a result, French learners of English will receive exercises for [p^h], and English learners of French will get exercises for [p].

Another example of these more detailed differences is that dental /t,d/ in Norwegian and the corresponding alveolar phones in English will trigger exercises for L2 learners across those languages, since these are no longer conflated into a single category. The same is true for other categories that were conflated in section 2.1, like the different versions of /r/ in French and English. Clearly, a phonetically correct realization of the rhotic in L2 will lead to a pronunciation which reduces the learner's foreign accent. This will make the sound easier to process for a native speaker, and less often lead to negative judgments of the speaker or to prejudices with regard to the learner's background. Cross-language similarity is discrete, and can only be defined continuously through perception tests.

3. Unresolved problems

While CALST helps L2 learners to focus their attention on new L2 sounds which diverge in their phonetic realization from the sounds in L1, it is far from perfect. In particular, as we pointed out in Section 1.2, the SLM states that the sounds to be learned should be defined at the level of positional allophones. While the multilingual phonetic categories used in L1-L2*map* and UPSID reflect fine phonetic detail beyond a purely phonemic level, they are not allophones. The lack of access to allophonic information in our database has a number of disadvantages.

The information in L1-L2*map* has been extended with positional information for a small subset of the languages in the database. This extension highlights sounds as new if they occur in a position in L2 which is not allowed in L1. Despite a large number of consonants in Mandarin Chinese, for instance, only /n/ and /ŋ/ occur syllable-finally, so that all other consonants will be connected to exercises in syllable-final position, even though the sounds themselves are familiar for Chinese learners.

But this positional information only shows the positions in which a phoneme occurs, and does not reflect positional allophones. For example, the database contains only a single phoneme /l/ for English, although the language has two positional allophones, clear [l] and dark [ɫ]. The /l/ defined in the database is the same as the lateral in German, which does not have these different positional variants. German learners' attention is therefore not drawn to the positional variation in English, and an incentive to create a separate phonetic category for English dark [ɫ] must therefore come from observation of this phonetic variant outside the pronunciation training platform, for instance in conversations with native speakers. Speakers of Japanese, on the other hand, will receive exercises for /l/ in English, since there is no corresponding speech sound in Japanese. Because CALST offers exercises for unfamiliar consonants in both initial and final position in the word, Japanese learners can observe the different variants, although the allophonic difference is not made explicit in the exercises, which are presented as exercises for /l/.

In section 2.1 we explained that Bulgarian learners do not get exercises for the French plosives. This is also true when we use cross-language phones, since French /p,t,k/ also occur in Bulgarian. We assumed that a Bulgarian learner of French would automatically pick the non-palatalized version of the stops when speaking French. This is probably incorrect, and they are more likely to use palatalized stops before the vowel /y/, for instance, pronouncing French 'tu' as /tju/ [16]. CALST uses self-monitoring in its pronunciation exercises (comparing the learner's pronunciation with that of the tutor), but does not evaluate whether the learner has acquired the L2 sound correctly. This would require exercises which contrast L2 phones with similar phones in L1 in a direct comparison, which is outside the scope of CALST at present (and difficult at best in future).

Although it seems necessary to train Bulgarian learners of French to produce the correct allophone, other types of allophonic variation may stem from more general phonetic principles and may not require explicit training. To our current knowledge, it is an open question whether learners who have acquired aspirated plosives in L2, for instance, automatically produce these without aspiration at the beginning of unstressed syllables or after an onset /s/, as is the case in English.

4. Summary

Current models of L2 learning use L1-L2 phone similarity to predict how difficult it is to acquire new L2 segments. Two practical implementations of similarity are proposed which use available linguistic information. Despite their weaknesses, they present a practicable approach to multilingual L2 pronunciation training.

5. Acknowledgements

The author would like to thank Jørgen Bosoni for discussion, and Øyvind Bech for the implementation of similarity measures in L1-L2*map*. This research was supported through project funding from Kompetanse Norge.

6. References

[1] J. Koreman, P. Wik, O. Husby and E. Albertsen, "Universal contrastive analysis as a learning principle in CAPT," *Proceedings of the workshop on Speech and Language Technology in Education* (SLaTE), 2013, pp. 172-177 (see also https://www.ntnu.edu/isl/calst).

[2] C. T. Best and M. D. Tyler, "Non-native and second-language speech perception: commonalities and complementarities," in M.J. Munro & O.-S. Bohn (eds.), *Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production,* pp. 13-34. Amsterdam: John Benjamins, 2007.

[3] O.-S. Bohn, "Cross-language and second language speech perception" (chapter 10), in E.M. Fernandez & H.S. Cairns (eds.), *Handbook of Psycholinguistics.* New York: Wiley and sons, 2017.

[4] J. E. Flege, "Second language speech learning: Theory, findings, and problems" (chapter 8), in W. Strange (ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233-277). Timonium, MD: York Press, 1995.

[5] U. Weinreich, "On the description of phonic interference," *Word*, vol. 13, 1957, pp. 1-11.

[6] W. Strange, "Cross-language phonetic similarity of vowels," in M. Munro & O.-S. Bohn (eds.), *Second Language Speech Learning: The Role of Language Experience in Speech Perception and Production*. Amsterdam: John Benjamins, 2007, pp. 35-55.

[7] J. Koreman, Ø. Bech, O. Husby and P. Wik, "L1-L2*map:* a tool for multi-lingual contrastive analysis, *Proceedings of the 17th International Congress of Phonetic Sciences* (ICPhS2011), Hong Kong, 2011 (see also https://L1-L2*map*.hf.ntnu.no).

[8] I. Maddieson, *UPSID: UCLA phonological segment inventory database.* UCLA Phonetics Laboratory, Department of Linguistics, 1980.

[9] I. Maddieson, *Patterns of Sounds.* Cambridge: Cambridge University Press, 1984.

[10] I. Maddieson, S. Flavier, E. Marsico and F. Pellegrino, *LAPSyD: Lyon-Albuquerque Phonological Systems Databases, Version 1.0.* http://www.lapsyd.ddl.ish-lyon.cnrs.fr/lapsyd/, 2014-2016.

[11] R. Lado, *Linguistics across cultures: Applied linguistics for language teachers.* Ann Arbor: University of Michigan Press, 1957.

[12] International Phonetic Association, *The Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet.* Cambridge: Cambridge University Press, 1999.

[13] C. T. Best, "A direct-realist view of cross-language speech perception," in W. Strange (ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research.* Trimonium, MD: York Press, 1995.

[14] M. J. Munro and T.M. Derwing, "Processing time, accent and comprehensibility in the perception of native and foreign accented speech," *Language and Speech,* vol. 38, pp. 289-306, 1995.

[15] P. Ladefoged and I. Maddiesson, *The Sounds of the World's Languages.* Oxford: Blackwell Publishers, 1996, p. 3.

[16] B. Nikolov, "Étude de phonétique et phonologie contrastives (domaines français et bulgares)," *Annuaire de l'Université de Sofia*, Faculty of Western Philologies, vol. 2, 1971, pp.1-75.