Towards a token-by-token whole-spectrum approach to sound change using deep learning: A case study of Khmer coda palatalization

Sothornin Mam[1], Francesco Burroni[1,2], Sireemas Maspong[1,2]
[1]Center of Excellence in Southeast Asian Linguistics and Department of Linguistics,
Faculty of Arts, Chulalongkorn University, Thailand
[2]Spoken Language Processing Group, Institute for Phonetics and Speech Processing,
LMU Munich, Germany
6681006122@student.chula.ac.th,
{francesco.burroni, s.maspong}@phonetik.uni-muenchen.de

Abstract

In this paper, we present a token-by-token whole-spectrum approach using deep learning to investigate sound change, focusing on the understudied phenomenon of Khmer velar coda palatalization. By applying deep learning classification models to Mel spectrograms, our approach confirms that Khmer is undergoing velar palatalization. The model also reveals significant inter-speaker variation within the same linguistic community, with different speakers at different stages of the sound change. Additionally, our method, using Grad-CAM, identifies specific acoustic features associated with this phonological shift. Our findings highlight the potential of deep learning techniques to enhance our understanding of sound change.

1 Introduction

A fundamental area in the study of phonology and phonetics deals with the study of how the sound inventories of a language evolve over time. This continuous process of sound change represents one of the most pervasive and characterizing properties shared by all human languages and it has been investigated since the 18^{th} century (*cf.* Garrett and Johnson 2013 and references therein).

Sound changes are traditionally considered the endpoint of low-level phonetic changes that gradually diffuse through lexical items and a population of speakers until they affect the total number of phonological contrasts by changing their phonetic realization or by increasing (via splits) or decreasing their number (via mergers). An outstanding issue in linguistic theory remains developing viable models that can render justice to the complexity of this process (*cf.* Harrington et al. 2018 and references therein).

Two important prerequisites stand in the way of developing appropriate models of sound change.

First, sound changes often involve a variety of acoustic (and articulatory) dimensions that are relevant to the production and perception of speech. In other words, to appropriately characterize sound change, we must be able to probe, describe, and quantify variation beyond a small number of low-dimensional phonetic parameters that are often examined in experimental phonetic and phonological studies, e.g., duration, vowel formants, fundamental frequency *etc.* (for examples of this approach *cf.* Gubian et al. 2015, Puggaard-Rode 2022).

Second, given the increased attention paid in linguistic theory to exemplar and episodic models of lexical access and speech production/perception and their relationship to sound variation and change (Pierrehumbert et al. 2002, Johnson 2007, Goldrick and Cole 2023, Blevins and Wedel 2009), we need to develop models that enable us to quantify variation of interest on an episodic or token-by-token basis.

In this paper, we present an approach that offers promising solutions to tackle the two issues outlined above and, thus, can help in developing comprehensive descriptions and models of sound change. Specifically, we present a deep-learning approach that (i) enables us to quantify multidimensional phonetic and phonological variation relevant to sound change by applying deep-learning classification to (Mel)-spectrograms and (ii) allows us to quantify the degrees of change associated with individual exemplars of a phonological category.

We apply this method to the phenomenon of palatalization in Khmer (ISO-693-3; khm), an Austroasiatic language and the official language of Cambodia.

1.1 The case study: Khmer velar palatalization

Khmer has a phonological contrast between velar and palatal nasal and stop consonants in the onset position. However, the status of this contrast in coda position remains debatable. According to descriptions in Khmer grammar books and dictionaries (e.g., Huffman, 1970; Filippi and Vicheth,

2016), velar codas /k/ and /ŋ/ undergo palatalization following front vowels, such as /iː/, /eː/, /ei/, /ɛː/, and /ae/, and are subsequently realized as palatal consonants [c] and [ɲ]. Furthermore, Khmer orthography only attests non-palatal coda following long front vowels. This suggests that palatal codas were not originally present, but developed over time through diachronic palatalization of velar codas in this environment.

Khmer palatalization is of great interest for three reasons.

First, no experimental investigations of the phenomenon exist. This constitutes a noteworthy empirical gap, given that palatalization phenomena *following* front vowels are relatively rare (being mostly known from Germanic, *cf.* Hall 2022) compared to palatalization of consonants *preceding* front vowels.

Secondly, although this palatalization is often described in the literature as a completed sound change, anecdotal evidence suggests that Khmer speakers may not fully perceive a merger between palatal and velar sounds in this context. This raises the possibility that the change may, in fact, *not* be fully complete. Some speakers may produce fully velar consonants, while others produce fully palatal consonants. Additionally, speakers might produce fronted velar consonants due to co-articulation with preceding front vowels. Consequently, the status of this phenomenon—as either an ongoing or completed sound change—remains uncertain and requires further investigation.

Third, the distinction between velar and palatalized velars, is well-defined articulatory in terms of tongue body contact with the different points of the palate, yet, the acoustic manifestations of this articulatory underpinnings are notoriously elusive (Keating and Lahiri, 1993; Ladefoged and Maddieson, 1996; Ladefoged and Johnson, 2014).

1.2 Research questions

With the issues outlined above in mind, we investigate Khmer palatalization with a token-by-token whole-spectrum approach that leverages deep learning.

First, we trained convolutional neural network models to classify Mel spectrograms of phonologically contrastive velar and palatal nasals in non-front vowels environment.

Subsequently, the trained models were then used to investigate Khmer palatalization of velar consonants. Specifically, they were used to predict the probability that a certain velar token is realized as palatal in front-vowel environments. This approach allows us to situate individual tokens from individual speakers on a velar to palatal continuum based on the whole Mel spectrogram.

Equipped with these models, we investigated the following three research questions.

(i) Do we observe a degree of palatalization of velar stop after front vowels in Khmer as reported in grammar and dictionaries?

(ii) Do we observe complete neutralization of velars to palatals after front vowels in Khmer or do we observe a cline of realizations; possibly, differing across individuals within a community?

(iii) Finally, can an investigation of the inner workings of said models help to shed light on the spectral features that are likely to underlie the (eroding) distinction between velar and palatal stops in Khmer?

2 Methodology

2.1 Participants and data collection

The recordings were collected from five native speakers of Khmer: two from Phnom Penh and three from Takhmao, a city near Phnom Penh. There are two male and three female participants. Their ages are in the range of 20-30 years old ($\mu = 23.8, \sigma = 3.83$). Speakers from both cities speak the Phnom Penh variety reported to have final velar palatalization (Filippi and Vicheth, 2016). All participants were literate in Standard Khmer.

The target words consisted of monosyllabic or minor disyllabic words with final palatal and velar nasals, preceded by both front and non-front vowels. We divided the target words into two groups: one containing true velar and palatal nasals, and the other containing palatalized velar nasals.

For the true velar and palatal dataset, the target words included those with velar /ŋ/ and palatal /ɲ/ nasals following non-front vowels /a/, /iə/, /uə/, and /ou/. In this environment, velar consonants are not expected to undergo palatalization. We prioritized minimal pairs between velar and palatal codas. There were 16 unique target words (2 codas $\times$ 4 vowels $\times$ 2 unique words per template). Participants were asked to produce each target word 20 times, resulting in a total of 320 tokens per speaker.

For the palatalized velar dataset, the target words included those with a velar nasal following front vowels /iː/, /eː/, /ei/, /ɛː/, and /ae/, which are environments for velar palatalization. There were 10 unique target words (5 vowels × 2 unique words per template), with each word repeated 10 times, resulting in 100 tokens per speaker. An example of the different types of words used in the wordlist is shown in Table 1. All target words were embedded in a carrier sentence: [niʔ.jɨj tʰaː _____ tɒː tiət] "Speak the word _____. Next.".

Palatal	Velar	Palatalized
baɲ 'to shoot'	baŋ 'to cover'	wɛːŋ 'to be long'

Table 1: An example of the wordlist

For the true velar and palatal dataset, the target words were presented to participants embedded in a carrier sentence in Khmer orthography. Participants were instructed to read the entire sentence aloud. For the palatalized velar dataset, to avoid the influence of the orthography on the final consonant, we included trials where participants were presented with pictures representing the target words, in addition to the trials with orthographic presentation. It is worth noting that we did not observe any difference between picture and orthography stimuli. Thus, we analyzed the tokens from both picture and orthography stimuli together in this paper.

The recordings were conducted using the SpeechRecorder software (Draxler and Jänsch, 2004). The audio signal was captured directly to a laptop computer at a sampling rate of 44.1 kHz through a head-mounted unidirectional microphone. The recordings were done in a closed space with minimal noise.

2.2 Data preparation

The recordings were force-aligned using the MAUS language-independent model (Schiel, 1999). Subsequently, the TextGrids generated by MAUS were manually corrected using Praat (Boersma and Weenink, 2020) by a phonetically trained native speaker of Khmer. The manual correction focused on the segmentation of the nasal finals to ensure that no trace of the vowel was included in the nasal coda segment, as the acoustic signals during the nasal closures were used as input for the classification model. An example of the segmentation is illustrated in Figure 1.

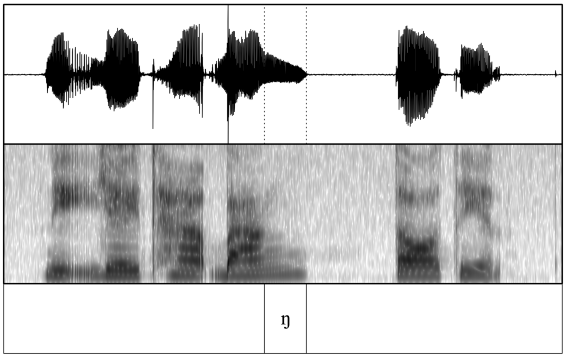

Figure 1: A segmentation example from Praat of word with velar nasal coda

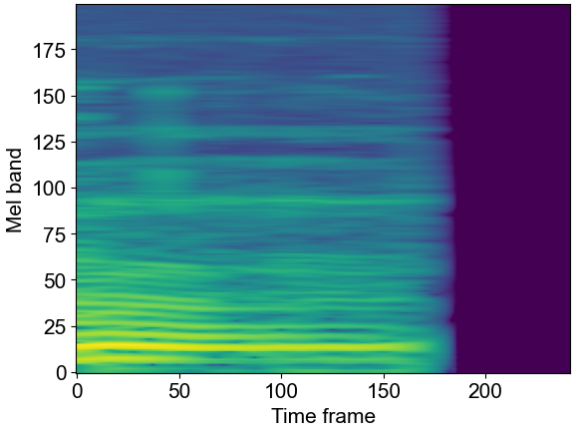

Figure 2: Mel-spectrogram example of one nasal token input. The dark shade part exhibits the zero-padded region.

To capture the multidimensionality of the acoustic signal, we extracted spectral information from the audio signal during the nasal closure using the Mel spectrogram method. The window size was set to 50 ms with a 1 ms time step, and the Mel filter bank was set to 200 Mel bands, ranging from a minimum frequency of 1 Hz to a maximum frequency of 22.05 kHz. To create a consistent input size required by the model, all tokens shorter than the maximum duration were symmetrically padded with zeros preceding and following the audio signal. Figure 2 illustrates an example of the resulting Mel spectrograms, showcasing the intensity of spectral components over time and frequency. The resulting Mel spectrograms were used as the input to the classification models.

2.3 Baseline model training and testings

To account for inter-speaker variation, separate deep learning models were trained using each individual participant's data, following the method-

ology of Liu and Xu (2023). As a result, we developed five baseline models, corresponding to the number of participants recorded for this study.

Each model is a deep learning classification model to classify the place of articulation of the coda based on spectral information. A convolutional neural network (CNN)-based model was implemented, with the best-performing model used as the baseline model. The model architecture is adapted from Doshi (2021) and its schematization is illustrated in Figure 3. The architecture consisted of four convolutional layers.

The input for the baseline model was the Mel spectrograms from the true velar and palatal datasets. The dataset was split into training, validation, and test sets in a 40:30:30 proportion, resulting in 128:96:96 tokens per model.

The model was trained using PyTorch in Python, with the Adam optimizer and a learning rate of 0.001. Binary cross-entropy loss was used to calculate the loss based on the probability values for both classes. The training process lasted for 150 epochs, and the best model was selected based on the lowest loss value on the validation set. The best-performing model was then used to classify the testing data to confirm its ability to accurately classify true velar and palatal codas.

In addition to the classification results, we also extracted a prediction probability representing the degree of palatalization on a scale from 0 (velar) to 1 (palatal). To achieve this, the sigmoid function was employed as the activation function applied to the output layer. The sigmoid function can be calculated using the following formula:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

The baseline model, trained on true velar and palatal codas, was then used to classify the palatalized velars based on their Mel spectrograms. The degree of palatalization was quantified by the sigmoid function as described above. We interpreted values closer to 0 as indicating that the palatalized velars are more velar-like, while values closer to 1 suggest that they are more palatal-like.

3 Results

3.1 Classification of true velar and palatal

The model trained on true velar and palatal codas successfully classified these true velar and palatal codas with 100% accuracy across all models for all participants. To further evaluate the performance of our model, we also extracted the probability outputs. The histograms of the probabilities from the model of all participants are illustrated in Figure 4.

The distributions for all speakers are clearly divided between the two classes. All tokens of each class were classified with extreme probabilities, either 0 or 1, with no tokens showing intermediate probability values. Specifically, all velar nasal tokens had values closer to zero, while palatal nasals had values closer to one. This provided strong evidence that the model effectively categorizes tokens based on their place of articulation and further confirmed that velar and palatal codas are contrastive in a non-front vowel environment.

3.2 Classification of palatalized velars

When the classification models were applied to palatalized velar tokens, two distinct patterns of classification emerged, as summarized in Table 2. Notably, there were no effects of gender or place of origin, Phnom Penh or Takhmao, on the pattern displayed by the speakers.

Participants	Palatalization pattern
SP3, SP4	Categorical
SP1, SP2, SP5	Gradient

Table 2: Summary speaker groups of the two types of sound change patterns.

For one group of speakers, SP3 and SP4, the models classified the majority of the palatalized velar tokens as palatal nasals /ɲ/ (> 90%), as shown in Table 3. This suggests that, for these speakers, velar nasals following front vowels undergo a categorical shift to palatal nasals. The histograms of the probability distribution for SP3 and SP4, shown in Figure 5, also reflect this categorical shift, with the majority of tokens clustering around the probability value of 1, indicating ubiquitous classification as palatals.

For the other group of speakers, SP1, SP2, and SP5, the models classified approximately half (50% - 70%) of the palatalized velars as palatals, with a slightly higher number of tokens categorized as palatals than velars. Notably, SP5 exhibited a larger proportion of palatal classifications compared to the other participants, with 67% of all tokens classified as palatal. In the histograms for SP1, SP2, and SP5, although two small peaks are observed at both

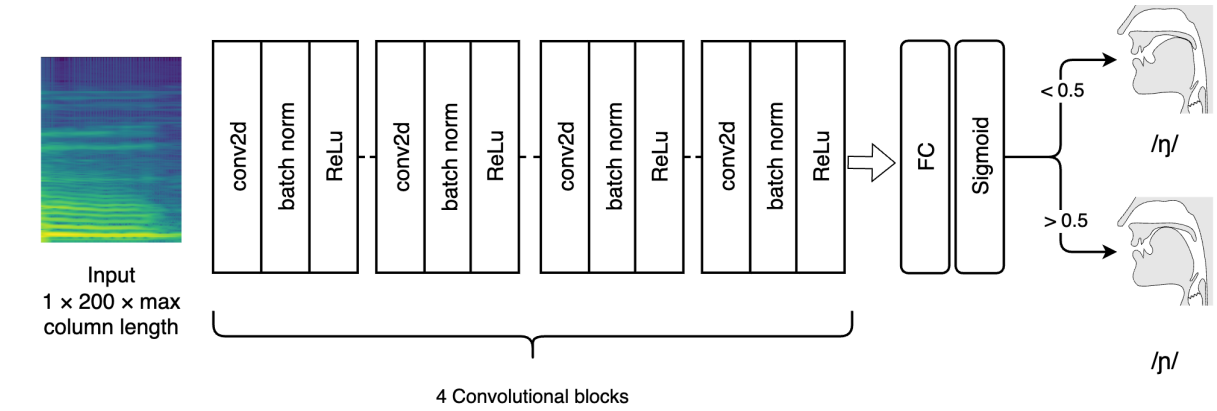

Figure 3: Audio classification model architecture.

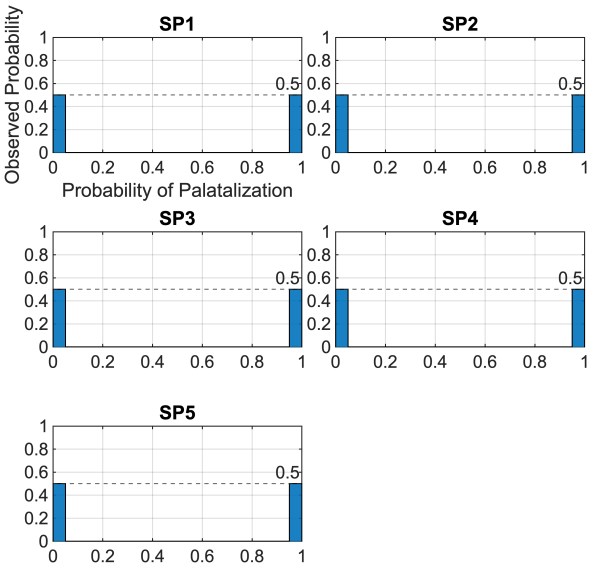

Figure 4: Histogram for true velar and palatal probability distribution of each participant.

ends of the distribution, the probability distribution is not as categorical. Mid-range probability values between 0 and 1 are sparsely distributed.

We may interpret these findings as suggesting that velar palatalization for this group of speakers is not a completely categorical process, but a gradient process. In other words, the contrast between velar and palatal nasals is not entirely neutralized in the front vowel environment: although some tokens may merge with true palatals, the majority of palatalized velars are not fully realized as palatals. These tokens might be realized as sounds intermediate between velars and palatals, likely due to the co-articulation effects of front vowels and velar codas where the tongue body position is intermediate between velar and palatal positions in the vocal tract.

Participant	/ŋ/	/ɲ/
SP1	47%	53%
SP2	47%	53%
SP3	7%	93%
SP4	2%	98%
SP5	33%	67%

Table 3: Palatalized velar class distribution of each participant.

3.3 Gradient-weighted Class Activation Mapping (GradCAM)

Given the model's strong performance in classifying true palatal and true velar nasals, this section investigates the acoustic features used by the models to distinguish between these two places of articulation. Previous acoustic studies have shown that this contrast is primarily signaled by differences in the transition of adjacent vowel formants, which are highly dependent on the vowel quality. For example, the formant transition from high front vowels to velar codas differs from that of low back vowels to velar codas (Ladefoged and Johnson, 2014). However, our findings in Section 3.1 demonstrate that the models accurately recognized the true place of articulation for the two nasal consonants using only the acoustic information from the nasal closure itself, without relying on vowel transitions. This suggests that the contrast between velars and palatals is also present within the acoustic signal of the consonants themselves.

To explore the spectral features that the models used to differentiate between the two places of articulation, we applied Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017) to the model classifications of true velars and true palatals as discussed in Section 3.1. Grad-

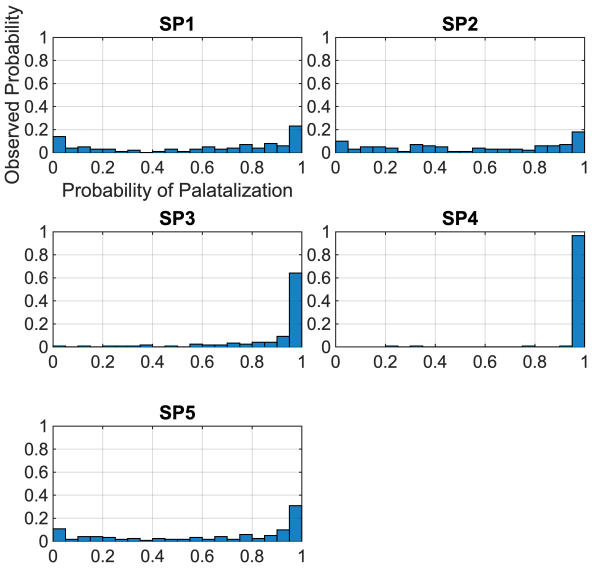

Figure 5: Histogram for palatalized velar probability distribution of each speaker.

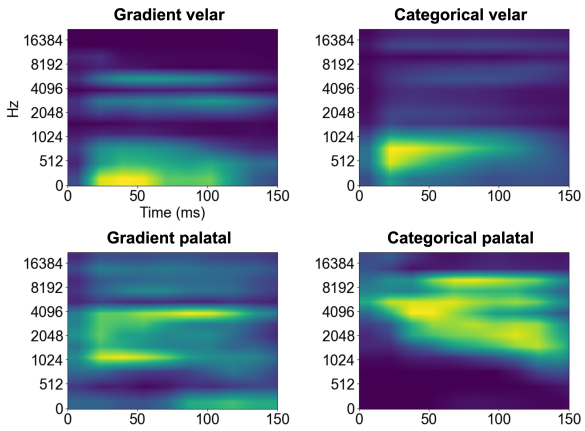

Figure 6: Grad-CAM class activation heat maps showing average weights for velar tokens (top) and palatal tokens (bottom) from SP2 exhibiting a gradient pattern (left) and SP4 exhibiting a categorical pattern (right).

CAM enables us to pinpoint specific regions within the input that the model focused on during its classification, thereby revealing the spectral features that distinguish velar and palatal nasals. In this section, we present results from two participants who exemplify the distinct realization patterns of palatalized velar outlined in Section 3.2.

The results of the Grad-CAM analysis are shown in Figure 6. The heat maps depict the average activation weights in the Mel spectrograms that the models used to classify true velars (top) and true palatals (bottom). Lighter colors indicate regions where the model assigned greater importance during classification. Notably, we observed several straight lines spanning the entire duration of the Mel spectrograms across all heat maps. This pattern suggests that the spectral features distinguishing velar and palatal nasals are consistent throughout the nasal closure interval, rather than being tied to specific temporal moments.

The distinguishing features appear to be on the spectral dimension. Specifically, it is likely that these straight lines on the heat maps represent anti-formants, with the distinguishing feature for velar and palatal nasals being the frequency ranges where the anti-formants are located.

For the velar tokens (top of Figure 6), the models focused on the lower frequency range. Although this pattern is consistent across both speakers, the specific frequency ranges where the model concentrated differ. For the speaker with the gradient distribution (SP2; top left of Figure 6), the model concentrated most heavily in the lower frequency range, particularly the 0-250 Hz range, with some areas of lower weights distributed in the 250-1024 Hz range, around 3000 Hz, and 6000 Hz. In comparison, for the speaker with the categorical distribution (SP4; top right of Figure 6), the model's focus was heaviest around 512 Hz, with additional areas of lower weights in regions comparable to those seen in the velar tokens produced by the speaker with the gradient distribution.

For the palatal tokens (bottom), the model's focus shifted to the higher frequency range. For the speaker with the gradient distribution (SP2; bottom left), the model concentrated the heaviest weight on two regions: around the 1024 Hz and 4096 Hz frequencies, with lower weights distributed across all frequency ranges. On the other hand, for the speaker with the categorical distribution (SP4; bottom right), the model's focus clustered in the higher frequency range, above 4000 Hz, with the heaviest weight lying between 8192 Hz and 16384 Hz.

There seems to be a tendency that if the main feature found for a velar is higher, as in the case of the speaker with categorical distribution, the corresponding feature for the palatal would also be higher.

In sum, based on the Grad-CAM results, we hypothesize that the distinction between velar and palatal nasals is present in the spectral domain, specifically in the location of anti-formants across different frequency ranges. However, there is still no clear evidence explaining why the two types of speakers, based on their production of palatalized

velars, differ in their production of true velars and true palatals as well. Further work is needed to fully elucidate this matter.

4 Discussion and Conclusion

Returning to our research questions, we first asked whether velar consonants in Khmer exhibit a certain degree of palatalization when they appear after front vowels, a process that has been reported as categorical in grammars and dictionaries. Our deep learning analyses confirm that a model trained on phonologically contrastive velar and palatal consonants classifies a large number of velar tokens as palatal in the environment following front vowels. This finding confirms an ongoing palatalization sound change in Khmer.

Additionally, we asked whether this sound change is completed and categorical or whether we observed a cline of velar to palatal realizations. To address this question, we have applied a method that allows us to quantify the degree of palatalization on token-by-token and subject-by-subject basis relying on the entire Mel spectrogram in view of known difficulties in characterizing place of articulation differences, especially for velar vs. palatal nasals. Our findings suggest that the process is an ongoing sound change as the realization of palatalized velars is not always identical to that of palatals. Interestingly, within the same speaker community, we observe that, for some speakers, the sound change is completed and palatalized velars are basically indistinguishable from phonological palatals. These findings resonate with the notion that sound change gradually diffuses through a community of speakers that propagate change via their interactions, in line with recent episodic approaches to sound change that rely on agent-based simulation (e.g., Harrington et al., 2018).

Finally, we also asked whether we can probe the inner workings of our model to relate our whole-spectrum analysis to low-dimensional phonetic features. Using Grad-CAM, we were able to hypothesize that the models are able to identify the place of articulation of Khmer consonants on the basis of a subset of frequency ranges in the spectrum that are broadly compatible with so-called anti-formants, as is known from the phonetics literature.

Beyond the empirical contribution of elucidating important details of a previously unstudied sound change, Khmer coda palatalization, we believe that this work also offers a first step towards an important methodological contribution. As noted in the introduction, recent works on sound change have emphasized the importance of the multidimensional richness of the acoustic signal and the role of episodic instantiations of these signals in influencing changes that affect the phonological categories of a language. In this paper, we have proposed a method that leverages deep learning and the entire Mel-spectrogram as a way to tackle these issues and quantify ongoing sound-change. The approach we have developed is able to probe, describe, and quantify the nature of a sound change and observe its distribution within a small sample of a linguistic community of interest. Additionally, we have offered preliminary ideas to bridge the gap between high-dimensional whole-spectrum analyses and more-traditional phonetic analyses.

The approach we have developed – we believe – is widely applicable to a variety of sound changes and is of interest to scholars working on the topic. This is because our approach offers a way to quantify where each episodic instantiation of a phonological category resides in a phonetic continuum between two phonological categories that may be drifting toward each other. This is of course a problem that is familiar from many types of sound changes, e.g., palatalization, lenition, changes in vowel quality *etc*. Our method, thus, constitutes an addition that can supplement the toolkit of experimental phonologists and phoneticians working on sound change.

To conclude, in this paper we have developed a token-by-token whole-spectrum approach that leverages deep learning. We have applied this method to a previously understudied case of sound change, Khmer coda palatalization. Our method has confirmed that the language is undergoing sound change, as hypothesized in previous work. Strikingly, different speakers within the same linguistic community seem to lie on different points along the path toward the completion of this sound change. Finally, we were able to relate the change in question to specific acoustic characteristics that are notoriously difficult to pinpoint. Thus, it is our hope that the findings and methods presented in this paper will offer a small further step towards a better understanding of a core property of human languages: the continuous evolution of their sound inventories.

References

Juliette Blevins and Andrew Wedel. 2009. Inhibited sound change: An evolutionary approach to lexical competition. *Diachronica*, 26(2):143–183.

Paul Boersma and David Weenink. 2020. Praat: doing phonetics by computer.

Ketan Doshi. 2021. Audio Deep Learning Made Simple: Sound Classification, step-by-step.

Christoph Draxler and Klaus Jänsch. 2004. SpeechRecorder – a Universal Platform Independent Multi-Channel Audio Recording Software. In *Proceedings of the fourth International Conference on Language Resources and Evaluation (LREC))*, pages 559–562, Lisbon.

Jean-Michel Filippi and Hiep Chan Vicheth. 2016. *Khmer pronouncing dictionary: Standard Khmer and Phnom Penh dialect*. UNESCO Office Phnom Penh/ KAM éditions, Phnom Penh.

Andrew Garrett and Keith Johnson. 2013. Phonetic bias in sound change. In *Origins of Sound Change: Approaches to Phonologization*. Oxford University Press.

Matthew Goldrick and Jennifer Cole. 2023. Advancement of phonetics in the 21st century: Exemplar models of speech production. *Journal of Phonetics*, 99:101254.

Michele Gubian, Francisco Torreira, and Lou Boves. 2015. Using functional data analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics*, 49:16–40.

Tracy Alan Hall. 2022. *Velar fronting in German dialects: A study in synchronic and diachronic phonology*. Language Science Press, Berlin.

Jonathan Harrington, Felicitas Kleber, Ulrich Reubold, Florian Schiel, and Mary Stevens. 2018. Linking cognitive and social aspects of sound change using agent-based modeling. *Topics in cognitive science*, 10(4):707–728.

Franklin E. Huffman. 1970. *Cambodian system of writing and beginning reader*. Yale University Press.

Keith Johnson. 2007. Decisions and mechanisms in exemplar-based phonology. *Experimental approaches to phonology*, pages 25–40.

Patricia Keating and Aditi Lahiri. 1993. Fronted Velars, Palatalized Velars, and Palatals. *Phonetica*, 50(2):73–101.

Peter Ladefoged and Keith Johnson. 2014. *A Course in Phonetics (seventh edition)*. Cengage Learning, Stamford.

Peter Ladefoged and Ian Maddieson. 1996. *The Sounds of the World's Languages*. Blackwell Oxford.

Zirui Liu and Yi Xu. 2023. Deep learning assessment of syllable affiliation of intervocalic consonants. *The Journal of the Acoustical Society of America*, 153(2):848–866.

Janet Pierrehumbert et al. 2002. Word-specific phonetics. *Laboratory phonology*, 7(1):101–140.

Rasmus Puggaard-Rode. 2022. Analyzing time-varying spectral characteristics of speech with function-on-scalar regression. *Journal of Phonetics*, 95:101191.

Florian Schiel. 1999. Automatic Phonetic Transcription of Non-Prompted Speech. In *Proceedings of the 14th International Congress of Phonetic Sciences*, pages 607–610.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.