INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

The Four-way Classification of Stops with Voicing and Aspiration for Non-native Speech Evaluation

Titas Chakraborty[1], *Vaishali Patil*[2], *Preeti Rao*[1]

[1]Department of Electrical Engineering, Indian Institute of Technology Bombay, India
[2]International Institute of Information Technology, Pune, India

titas0602@gmail.com, vvpatil2429@gmail.com, prao@ee.iitb.ac.in

Abstract

The four-way distinction of plosives in terms of voicing and aspiration is rare in the world's languages, but is an important characteristic of the Indo-Aryan language family. Both perception and production pose challenges to the language learner whose native tongue does not afford the specific distinctions. A study of the acoustic-phonetics of the sounds and their possible dependence on speaker characteristics, such as gender or native tongue, can inform methods for accurate feedback on the quality of the phones produced by a non-native learner. We present a system for the four-way classification of stops building on features previously proposed for aspiration detection in unvoiced and voiced plosives. Trained on an available dataset of Hindi speech by native speakers, the system works reliably on production data comprising Bangla words uttered by native Bangla and non-native (American English L1) speakers. The latter display a variety of articulation patterns for the given target contrasts, providing useful insights related to L1 influence on the voicing-aspiration production in word-initial CV contexts.

Index Terms: Voiced aspirates, four-way classification of plosives, pronunciation scoring, L1 influence

1. Introduction

Plosives, including stops, play a prominent role in the phonetic inventory of any language given the range of stable and perceptually distinct variations available within this category of consonants. A subset of the contrasts in the dimensions of place, voicing and manner of articulation are adopted in any given language. The manner of articulation includes the presence or absence of aspiration accompanying the plosive burst release. Aspiration of unvoiced plosives is allophonic in several languages appearing, for instance, in word-initial contexts. Hence, studies on phonemic aspiration have been mostly confined to unvoiced consonants with voice onset time (VOT), also known as the release duration, being the studied distinctive acoustic feature [1].

The Indo-Aryan language group is among the few language groups of the world where aspiration is a phonemic attribute in both unvoiced and voiced plosives, i.e. we see the 4-way contrast of plosives for each place of articulation [2]. Aspirated and unaspirated plosives are distinguished mainly by the aspiration phase that follows the burst release so that the aspiration is perceived as a release of breath accompanying the plosive. The acoustic cue (equivalent to the VOT) in the case of voiced plosives is the lag-VOT or vowel onset time. In the CV context, the aspiration (arising from increased glottal opening just preceding the plosive burst release) extends into the vowel region giving rise to breathy voice quality in the initial part of the vowel. Thus, apart from voice or vowel onset time, acoustic cues associated with the other articulatory dimensions of greater glottal opening and gradual closure as well as the co-articulatory breathy voice quality of the vowel play a role [3], [4], [5].

In previous work with Indo-Aryan languages, Davis [6] found that the lag-VOT discriminated all four velar stops in Hindi word-initial utterances. Dutta [7] conducted an acoustic-phonetic study of the four-way contrast of Hindi stops, with spectral analysis in the vowel region following the voiced aspirated stops indicating breathy characteristics over a substantial portion of the vowel. Mikuteit [8] investigated inter-vocalic geminate consonants of Bengali to find that the release duration or VOT was found to discriminate both the voiced and unvoiced consonants by aspiration. The beginning of the release was marked by the first burst at the end of the closure. The end of the release was marked where the first regular glottal pulses of the following vowel appeared in connection with a rising amplitude. It is noted that aspiration plays an important role in the pronunciation of Bangla words with the occurrence of several minimal word pairs[9].

In a study that covered both stops and affricates, word-initial plosives of Marathi were shown to separate better into aspirated and unaspirated classes when breathy vowel quality features were combined with the voicing onset time [10], [11]. Patil and Rao [12] proposed several acoustic measures to detect the aspiration contrast, separately for unvoiced and voiced plosives in CV contexts in Marathi. While the VOT separated the unvoiced consonants, features based on voice quality were necessary for reliable discrimination. Detection of aspiration for the voiced plosives was more challenging with an achieved accuracy of 85 % versus the 90 % for unvoiced aspiration detection on a 20-speaker dataset of word-initial plosives in all 8 vowel contexts The same GMM likelihood classifier trained on the Marathi dataset was found to perform with similar accuracy on Hindi speech by native Hindi speakers, attesting to the robustness of the acoustic-phonetic features across speakers and language contexts. A comparison with an automatic speech recognition system using standard spectral features (MFCC) demonstrated a superior performance for the acoustic-phonetic classifier in both within language and cross-language training-testing scenarios [12]. In this paper, we extend this previous work to the joint classification of voicing and aspiration motivated by a pronunciation scoring application that requires this discrimination. While retaining a subset of the most reliable aspiration detection features, we investigate further the voiced-unvoiced distinction across aspiration classes.

The 4-way contrast of plosives, where voicing and aspiration are distinctive for a given place of articulation, is challenging to acquire for a new learner whose native language (L1) does not possess the phonemic aspiration contrast. While the learner typically replaces the unfamiliar phone with the nearest available L1 phone, a high degree of variability can be expected in the acoustic realization especially if the distinction encom-

http://dx.doi.org/10.21437/Interspeech.2021-635

passes more than one articulatory dimension. Automatic detection of the nature of of deviation can be very useful as feedback in language learning contexts. It can potentially also aid neuroscientific studies focusing on representations of the distinctive features of speech in the auditory cortex [13].

In the current work, we consider a task relevant to the acquisition of the 4-way plosive contrast by speakers of American English (AE) as L1 in the word-initial CV context. The target language is Bangla which differs from English in a number of segmental (phone classes) as well as suprasegmental (e.g. lexical stress) aspects. In several 2-way contrast languages such as English, the so-called voiced plosives in word-initial position are phonetically unvoiced unaspirated, whereas the unvoiced plosives are phonetically unvoiced aspirated [14]. The goal of this work is to propose acoustic features that reliably capture the production characteristics of non-native learners of the four-way contrast of plosives characteristic of Indo-Aryan languages. With a suitable dataset of Hindi speech already available, we use the acoustic features previously demonstrated for Hindi aspiration detection while augmenting these with features for voicing. A random forest ensemble classifier is trained on the Hindi speech dataset. The performance of the automatic classification is reported on a dataset of native Bangla speech and on recordings of non-native (AE L1) speakers uttering the Bangla words. In the next section, we review the relevant language characteristics via a description of the datasets used in this work. This is followed by a discussion of the signal analysis applied to word utterances to obtain the acoustic features for the four-way classification. Classification experiments are presented in Section 4, and the results discussed.

2. Datasets

An available Hindi speech dataset first presented by Patil and Rao [12], was used in this work for the parameter tuning of the acoustic-phonetic feature extraction methods. The new datasets created for this work are smaller in the number of speakers due to limitations from the pandemic lockdown. All audio is sampled at 16 kHz.

Dataset A: The native Hindi speech dataset comprises segmented words by 20 speakers (10 male, 10 female) uttering the Hindi words in 2 different sentence contexts with 8 distinct words per stop phone. The words were presented as text prompts to be read aloud by the speaker. Each of the words contains the specific stop phone in word-initial position followed by each of the 8 distinct vowels of Hindi in order to cover all the expected coarticulatory contexts for the stops. We have 4 stop phones per PoA (except for the labial stop which does not have the unvoiced unaspirated phone). With 4 distinct PoAs, we get in all 15 distinct stop phones. The word list in IPA notation, arranged by voicing and place of articulation, is provided by Patil and Rao [12], Appendix A; we omit the 4 affricate subgroups. We thus have a total of 15x8x2 tokens as the number of word utterances by each native speaker for a total of 4800 (20x15x16) utterances across speakers.

Dataset B: Ten native Bangla speakers (8 male, 2 female) read aloud 16 distinct Bangla words, one for each of the plosives listed in Table 1. The word-initial stops are drawn from the set of plosives with 4 places of articulation with an 2 repetitions of each word per speaker (except for 2 speakers who had 1 instance of each word and a third who had 28 usable utterances) for a total of 284 utterances. Five of the 10 speakers submitted recordings carried out remotely on their personal devices necessitating some noise suppression before further processing.

Dataset C: We had four non-native speakers (3 female, 1 male), each uttering between 8 to 14 words for a total of 50 utterances across the words listed in Table 1. The data of the AE L1 speakers, with no prior exposure to Bangla, was recorded in a "listen and repeat" task in the context of an auditory neuroscience research pilot at the Georgetown University Medical School. A native speech audio recording of the word was played out as the prompt to the non-native speaker with instructions to repeat what they perceived as closely as possible; no feedback was provided to the speakers on the quality of their production at any time. The obtained utterances were later labeled via a perception test involving 3 native listeners. These listeners had no access to the word intended by the speaker (i.e. the target) but could listen to a token repeatedly. Each of them labeled a non-native utterance using one of 5 labels corresponding to the 4 voicing-aspiration conditions and a 'none' label to capture what they perceived as ambiguous or invalid pronunciation. Each utterance was then assigned a 'perceived' label (to be treated as the ground truth for the evaluation of the automatic classifier) based on the majority vote, if there was one, else it was labeled 'none'. The native listeners were also previously administered randomly drawn tokens across all the distinct words from the native Bangla speech dataset to confirm that they could rate the native speech with perfect accuracy.

Table 1: *The list of Bangla words used in Datasets B and C with corresponding word-initial plosive attributes*

Place of Articulation	Voicing and Aspiration	Bangla Script	Word (IPA)
Velar	Unvoiced-Unaspirated	ক	kala
	Unvoiced-Aspirated	খ	kʰala
	Voiced-Unaspirated	গ	ɡola
	Voiced-Aspirated	ঘ	ɡʰola
Retroflex	Unvoiced-Unaspirated	ট	ʈala
	Unvoiced-Aspirated	ঠ	ʈʰala
	Voiced-Unaspirated	ড	ɖali
	Voiced-Aspirated	ঢ	ɖʰali
Dental	Unvoiced-Unaspirated	ত	t̪ika
	Unvoiced-Aspirated	থ	t̪ʰika
	Voiced-Unaspirated	দ	d̪ola
	Voiced-Aspirated	ধ	d̪ʰola
Bilabial	Unvoiced-Unaspirated	প	piʈa
	Unvoiced-Aspirated	ফ	pʰiʈa
	Voiced-Unaspirated	ব	baʃa
	Voiced-Aspirated	ভ	bʰaʃa

3. Methods

The processing of the recorded word utterances involves the stages of phone segmentation (or acoustic landmark detection) and the computation of suitable spectral and temporal features around the detected landmarks of burst release and vowel onset. Based on our previous work, reviewed in Section 1, we base the detection of aspiration on the acoustic cues of release duration and the noisiness in the vowel region. The voicing status is inferred from the automatic detection of the presence or absence of the voice bar preceding the burst release.

3.1. Landmark detection

In general, an analysis with high temporal resolution is necessary over the CV segment in order to locate the landmarks with sufficient precision. Patil and Rao [12] achieve this by first coarsely locating phone boundaries via forced alignment with the known transcript of the utterance using an available trained ASR. To accommodate possible articulation errors by the speaker, the ASR models corresponded to broad phone classes such as silence, obstruents and vowels trained on a large labeled dataset of continuous speech in the language of interest. In the present work, we opt to replace this step with signal processing approaches that do not rely on language-specific training and are therefore expected to be more robust across the language and accent variations of interest to us. Given that we want to identify the initial CV region in the isolated word segment, we employ a syllable detection algorithm to find the first vowel nucleus[15]. Peak-picking is applied to band-weighted short-time energies at 10 ms resolution with voicing strength used further to eliminate any local peaks around the plosive burst. The signal interval prior to the earliest detected vowel peak in the word utterance is then searched with the fine time resolution of 1 ms for the two acoustic landmarks of burst onset and vowel onset point as follows.

The onset of the burst is marked by the rapid rise in the short-time energy in the high frequency region (3500 to 8000 Hz) at the end of the preceding closure. It is detected via the peak in the corresponding rate-of-energy-rise contour computed every 1 ms with the energy over a 4 ms window. Next, the transition into the vowel is marked by a drop in the energy of the linear prediction residual signal relative to the speech signal energy. The residual of pre-emphasized speech is computed using LP analysis of order 16 (to model short-term spectral correlation in the 8 kHz bandwidth) over 25 ms windows with 10 ms shifts. The vowel onset is detected from the change in the ratio of short-time signal energy of the speech signal to that of its LP residual[16]. The vowel onset point is marked at the first detected glottal epoch where the ratio crosses a predetermined threshold as long as the signal energy at this point is high enough compared to the maximum signal energy attained in the vowel region. In the present work, we do not have manually labelled landmarks; we tune the critical parameters, viz. the thresholds in syllable peak picking and vowel onset point detection by maximising the 4-way classification accuracy on the native speaker dataset.

3.2. Feature extraction

We target the 4-way classification of stops in the same word-initial contexts as opposed to the previous 2-way classification task for each known voicing condition [12]. Voiced plosives are characterised by glottal vibration throughout the closure and release interval. The voice bar, arising from the sound radiated through the laryngeal wall during closure, appears as a very low-frequency periodic signal. The release duration (VOT) was found to clearly discriminate unvoiced aspirated plosives from the three remaining classes in Hindi and Marathi word-initial contexts [12]. The reliable detection of aspiration in voiced plosives, on the other hand, required the further analysis of the initial part of the vowel following the plosive for breathiness. We therefore use the following three acoustic features to achieve the four-way classification of stops across places of articulation as presented here. The resulting 3-dimensional feature vector obtained for a given word utterance is used as training or test data in a tree-based ensemble classifier.

(i) Vowel onset time (VOT) is computed as the time interval between the detected landmarks of the burst onset and the vowel onset. We note that the reliability of this parameter depends on the accuracy of the previous landmark detection.

(ii) Voicing is detected by looking for the presence of the voice bar in the closure duration. The signal when inverse filtered with an LP filter yields a very low energy residual for voiced stops that is captured well by the ratio of the short-time signal energy to residual energy (similar to that computed post burst release to identify the vowel onset point) [17]. The computed voicing feature corresponds to the ratio averaged over 3 windows of 25 ms windows separated by 10 ms shifts and takes on high values when the voice bar is present.

(iii) The ratio of the energy in the harmonics to that of the noise (termed SNR here) serves as a measure of the strength of the aspiration noise that accompanies glottal vibration post the vowel onset. A noise floor, corresponding to between-harmonics spectral power, is estimated using cepstral filtering [18]. The SNR feature is the ratio of speech signal power to this estimated aspiration noise power. It is computed over windows of 20 ms with 1 ms shift over a region of the signal following the detected vowel onset landmark. The region of analysis is critical given that different languages have been observed to show different extents of breathiness extending into the vowel [19]. Given that vowel onset detection accuracy is not perfect, especially in the case of the voiced stops, we anchor the analysis region to the previously detected syllable nucleus midpoint and use the maximum value estimated over 15 adjacent windows terminating at this vowel peak.

3.3. Classification

The extracted features are used with a random forest ensemble classifier due to its competitiveness with the best available models given modest sized datasets [20]. We fix the number of estimators to 300 and use all the three features for splitting at nodes. The maximum depth is retained as a tunable parameter in the range 3 to 8, with lower depths helping to avoid overfitting to training data while not compromising accuracy.

4. Experimental Results and Discussion

We first present the four-way classification performance achieved in leave-one-speaker-out mode on the Hindi speech dataset, which we also use to tune the landmark detection algorithm parameters discussed in Section 3.1. The selected classifier tree depth is 4 since higher depths change performance only marginally. The classifier is then trained on the entire 20-speaker Hindi dataset and its performance evaluated on each of the native Bangla and non-native speech datasets. In the case of the non-native dataset, it is also of interest to estimate the ability of the system to detect misarticulated words (i.e. deviations from the expected word) in terms of recall and precision.

4.1. Hindi speech dataset

The 4-way classification performance aggregated across the 20 speakers, where each test speaker data is held out from the training data, is presented in the confusion matrix of Table 2 where the labels are unvoiced-unaspirated (UV), unvoiced-aspirated (UVA), voiced-unaspirated (V) and voiced-aspirated (VA). The tunable parameters for landmark detection are varied over predecided ranges to maximize the overall 20-fold CV classification accuracy. We obtain a mean accuracy of 80.2% (standard deviation = 6%) over the 4800 utterances by 20 speakers.

We observe from Table 2 that the VA class has the lowest detection accuracy with most confusions arising in the aspiration attribute. The unvoiced utterances demonstrate much better aspiration discrimination. The observations are consistent with those of Patil and Rao [12] who attribute this to the more challenging vowel onset detection task in the case of the voiced stops. We see that voicing, on the other hand, is very well discriminated with a 2-way classification accuracy of 97.0% across aspirated and unaspirated stops.

Table 2: *Confusion matrix of model predictions for native Hindi utterances in cross-validation mode (values represent % of corresponding number of target words)*

	Model prediction			
Target phone	UV	UVA	V	VA
UV	83.5	13.6	2.8	0.1
UVA	12.4	86.5	0.6	0.5
V	0.9	2.6	85.2	11.3
VA	1.4	2.7	28.8	67.1

To examine gender dependence, if any, we report accuracies separately for male and female speakers (each obtained in 10-fold CV mode). We obtain the close performances of 81.4% and 79.1% for male and female speakers respectively.

4.2. Native Bangla speech dataset

Table 3 presents the confusion matrix obtained across speakers on the native Bangla utterances using the model trained on the Hindi speech dataset. An overall 4-way classification accuracy of 86.6% is achieved. We observe that both voicing and aspiration are discriminated with accuracies that exceed those reported on the Hindi speech dataset. While this seems puzzling, a probable explanation is the choice of words in the Bangla task. Unlike the diverse coarticulatory contexts of the Hindi target words that systematically covered all the vowels of the language, the Bangla targets are restricted to heavy syllables with the word-initial CV containing a long vowel only. Coupled with the fact that the Bangla speakers uttered each word in isolation, we expect relatively clearly articulated phones for the word initial CV compared to those of the Hindi speech dataset. The VA phones are particularly well identified by the model. We note a lower voiced stop detection accuracy which is found to arise from several tokens of one particular speaker falsely classifying as unvoiced, partly attributable to the background noise suppression that was required for the speaker's recording.

Table 3: *Confusion matrix of model predictions for native Bangla utterances (values represent % of corresponding number of target words)*

	Model prediction			
Target phone	UV	UVA	V	VA
UV	90.3	8.3	0	1.4
UVA	11.1	88.9	0	0
V	7.1	2.9	77.1	12.9
VA	1.4	4.3	4.3	90

4.3. Non-native speakers dataset

Table 4 presents a confusion matrix that captures the non-native articulations of the target phones as perceived by native listeners (majority vote) in the perception test. All the off-diagonal entries and the column marked 'N' represent incorrect utterances. We observe that while the UVA phones are realised correctly, the UV phone targets are realised as UVA or as ambiguous. This can be explained by the allophonic use of aspiration in word-initial contexts in the speakers' L1. English speakers discriminate unvoiced from voiced word-initial stops with aspiration. This also leads to occassional devoicing of voiced stops as see in row 3 of Table 4. As expected, the VA stops are seen to be always mispronounced but with different alternate realizations. Most common is more familiar unvoiced aspirated stop in word-initial position. In the cases of replacement with voiced stops, we noted that the speaker registered only a shift in stress placement. That is, the word was uttered with lexical stress on the second syllable for the voiced target and on the first syllable in the case of the voiced aspirated target. Bangla, like other Indo-Aryan languages (and unlike AE), is syllable-timed and does not specify lexical stress. Finally, we noted that many of the invalid utterances of VA phones demonstrated an onset of aspiration that was perceptibly delayed with respect to the burst release (rather than synchronous with it), pointing to the complexity of the required articulatory gesture.

Table 5 provides the confusion matrix for the automatically predicted class for the given perceived classes. We see that both UV and UVA utterances are correctly classified by the system. The system makes voicing detection errors in the case of voiced stop realizations. We do not have an 'N' at the classifier output as yet; this can be incorporated in future based on the class confidences tuned on a larger non-native dataset. Given our interest in the automatic detection of mispronunciations as well as specific articulation errors for feedback, we report two performance measures on the non-native dataset. The detection of mispronunciations is reported in terms of recall (% of mispronunciations correctly detected) and precision (% mispronunciation detections that are correct with reference to the corresponding ground-truth). From Table 5, we obtain a high recall of 92% and precision of 82%. The 4-way classification accuracy is 68% (getting to 83% if the tokens labelled 'N' are left out).

Table 4: *Perceived phone versus the target phone from the listening test on non-native (AE L1) Bangla word utterances*

	Perceived phone				
Target phone	UV	UVA	V	VA	N
UV	5	5	0	0	2
UVA	1	10	0	0	0
V	2	0	10	0	1
VA	0	6	2	0	6

Table 5: *Model prediction versus perceived phone for non-native (AE L1) utterances of the Bangla words*

	Model prediction			
Perceived phone	UV	UVA	V	VA
UV	7	1	0	0
UVA	0	21	0	0
V	6	0	6	0
VA	0	0	0	0
N	5	3	1	0

5. Conclusion

A native-speech trained classifier based on the acoustic-phonetic features of the four-way contrasted stops showed high accuracy in detecting articulaton errors by non-native learners. Our analysis reveals interesting insights about the diverse range of non-native realizations, explainable by the differences between the target phone and nearest L1 phone including the factors of allophonic usage and lexical stress. This underlines the potential for focused feedback based on future training with a larger dataset of perceptually labeled non-native speech.

6. Acknowledgements

The authors thank Dr. Max Riesenhuber for the non-native speech dataset used in this work.

7. References

[1] L. Lisker and A. Abramson, "Cross-language study of voicing in initial stops: Acoustical measurements," *Word*, vol. 20, pp. 384–422, 1964.

[2] P. Bhaskararao, "Salient phonetic features of indian languages," *Sadhana*, vol. 36, pp. 587–599, 2011.

[3] S. Mikuteit and H. Reetz, "Caught in the act: The timing of aspiration and voicing in east bengali," *Language and Speech*, vol. 50(2), pp. 247–277, 2007.

[4] C. Ishi, "A new acoustic measure for aspiration noise detection," *Proceedings of International Conference on Spoken Language Processing, Jeju Island, Korea*, pp. 629–634, Aug. 2004.

[5] H. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *Journal of the Acoustical Society of America*, vol. 101, pp. 466–481, 1997.

[6] K. Davis, "Stop voicing in hindi," *Journal of Phonetics*, vol. 22, pp. 177–193, 1994.

[7] I. Dutta, "Four-way contrast in hindi: An acoustic study of voicing, fundamental frequency and spectral tilt (Ph.D. dissertation). University of Illinois at Urbana-Champaign," 2007.

[8] S. Mikuteit, "Voice and aspiration in german and east bengali stops: A cross-language study," *Interspeech*, pp. 2873–2876, Sep. 2005.

[9] B. Barman, "Distinctiveness of aspiration in bangla," *Daffodil International University Journal of Business and Economics*, vol. 3, pp. 191–203, 2008.

[10] V. Patil and P. Rao, "Acoustic features for detection of aspirated stops," *In Proceedings of National Conference on Communication. Bangalore, India*, pp. 1–5, Jan. 2011.

[11] V. Patil, P. Rao, "Acoustic features for detection of phonemic aspiration in voiced plosives," *In Proceedings of Interspeech, Lyon, France*, pp. 1761–1765, Aug. 2013.

[12] V. Patil and P. Rao, "Detection of phonemic aspiration for spoken hindi pronunciation evaluation," *Journal of Phonetics*, vol. 54, pp. 202–221, Jan. 2016.

[13] J. Rauschecker and S. Scott, "Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing." *Nature Neuroscience*, vol. 12(6), pp. 718–724, 2009.

[14] R. Dixit, "Glottal gestures in hindi plosives," *Journal of Phonetics*, vol. 17, no. 3, pp. 213–237, Jul. 1989.

[15] K. Sabu, P. Rao, S. Chaudhari, and M. Patil, "An optimised signal processing pipeline for syllable detection and speech rate estimation," *Accepted for National Conference on Communications, Kharagpur, India (https://arxiv.org/abs/2103.04346)*, Feb. 2020.

[16] S. R. M. Prasanna and B. Yegnanarayana, "Detection of vowel onset point events using excitation information," *In Proceedings of Interspeech*, pp. 1133–1136, 2005.

[17] N. Dhananjaya, S. Rajendran, and B. Yegnanarayana, "Features for automatic detection of voice bars in continuous speech." *Interspeech*, pp. 1321–1324, 2008.

[18] P. J. Murphy and O. O.Akande, "Noise estimation in voice signals using short-term cepstral analysis," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1679–1690, 2007.

[19] P. Ladefoged and I. Maddieson, *The sounds of world's languages*. Blackwell Publishing, 2005.

[20] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *Journal of Machine Learning Research 15*, pp. 3133–3181, 2014.