Audio Engineering Society

Convention Paper 10529

Presented at the 151st Convention
2021 October, Online

Phoneme Mappings for Online Vocal Percussion Transcription

Alejandro Delgado[1,2], Charalampos Saitis[2], and Mark Sandler[2]

[1]*Roli Ltd., London, England, UK*
[2]*Queen Mary University of London, London, England, UK*

Correspondence should be addressed to Alejandro Delgado (alejandro@roli.com)

ABSTRACT

Vocal Percussion Transcription (VPT) aims at detecting vocal percussion sound events in a beatboxing performance and classifying them into the correct drum instrument class (kick, snare, or hi-hat). To do this in an online (real-time) setting, however, algorithms are forced to classify these events within just a few milliseconds after they are detected. The purpose of this study was to investigate which phoneme-to-instrument mappings are the most robust for online transcription purposes. We used three different evaluation criteria to base our decision upon: frequency of use of phonemes among different performers, spectral similarity to reference drum sounds, and classification separability. With these criteria applied, the recommended mappings would potentially feel natural for performers to articulate while enabling the classification algorithms to achieve the best performance possible. Given the final results, we provided a detailed discussion on which phonemes to choose given different contexts and applications.

1 Introduction

Music Information Retrieval (MIR) has had a growing influence on the music industry in recent decades. Results in tasks like music genre recognition, chord estimation, and source separation [1] bring optimism to the field in this regard.

Query By Vocal Percussion (QVP) is a field within MIR that uses percussive vocal sound events to retrieve information, usually other percussion sounds. These vocal percussion sound events are articulated so as to communicate a rhythmic idea, usually by imitating the sound of percussive instruments like those featured in a drum set. As such, the sounds and their dynamics could be transcribed so as to create realistic drum loops in seconds, making composers save time and effort prototyping rhythms without actual music knowledge. This is done through a process known as *Vocal Percussion Transcription* (VPT). In it, vocal percussion sound events are detected, analysed, and mapped to particular events (e.g. drum sounds) that compose a transcription file. For instance, if the system detects a vocal percussion sound event that is supposed to trigger a snare drum sound, the system would automatically annotate a snare drum symbol in that precise time frame within the transcription file.

A VPT system has the potential to operate offline or online (real-time). In offline mode, the algorithm has access to the whole audio file containing the vocal percussion performance and can use all that information at once to detect the onsets of vocal percussion sound events and classify them. The system would output a transcription file afterward with the sound events and

their timings. Conversely, in online mode, the algorithm has access to a short analysis buffer that contains the most recent few milliseconds of the recorded audio stream. In this case, the system would have to detect, classify, and usually trigger the response very shortly after the sound event is recorded; for instance, it would trigger a snare drum sound almost at the same time as the performer vocalises the percussive sound event that is supposed to trigger it. This online procedure puts an important constrain to the system, forcing a trade-off between delay (length of the analysis buffer) and performance (detection and classification accuracy). In this sense, the longer the analysis buffer, the more information is available to the algorithm and the better the performance is expected; but also the more delay between the trigger and the response, which could be perceptually unpleasant if it exceeds a certain threshold that usually depends on the nature of the task at hand.

Vocal percussion sound events are mostly composed of plosive, fricative, and affricative phonemes. As the delay buffer to trigger percussive responses is generally short, some of these phonemes are more distinguishable than others within their first few milliseconds. For instance, the phonemes /p/ and /s/ are vocally articulated in a very different way and therefore are likely to be easily discernible, while the phonemes /p/ and /b/ are articulated in a very similar way and are usually harder to separate, even when having access to the complete sound events. Also, for the process to feel natural to performers (usually music producers and musicians) it is convenient for them to vocalise some phonemes and not others on the basis of their resemblance with their response triggers. For example, if the response trigger is a kick drum sample, one would prefer the phoneme /p/ to the phoneme /s/, as the former's sound reminds more of a generic kick drum than the latter.

The motivation for this study is urged by these two previous facts. More specifically, we try to answer the following question: which group of phonemes provide optimal online VPT performance while remaining natural for the performer to use?

An essential first step to find the best set of phonemes to query drum sounds in real-time was to find a dataset of vocal percussion sound events that had phoneme annotations. For this, we chose the Amateur Vocal Percussion (AVP) dataset [2] as our reference dataset. This publicly available dataset contains vocal percussion recordings from 28 people with little or no experience in beatboxing (i.e., amateur participants) with a total of 4873 vocal percussion sound events (personal subset) imitating four drum instrument types: kick drum, snare drum, closed hi-hat, and opened hi-hat. Every vocal percussion sound event in the dataset has annotated onset, instrument label, and phonetic representation. In the case of the reference drum sounds, we took samples from InMusic's BFD3 library [3] at velocities 64 and 127. These included 150 kick drum samples, 274 snare drum samples, 276 closed hi-hat samples, and 276 opened hi-hat samples. Like the sound events relative to phonemes in the AVP dataset, these were sampled at 44100 Hz, and we applied dither to downscale their bit-depth from 24 to 16 bits so that it matched the one from phoneme sounds.

Once we had the data, we evaluated the appropriateness of each phoneme to make the final set on the basis of three criteria. The first criterion was the *frequency of use* of the phonemes, i.e., how many participants chose these phoneme-to-instrument mappings. The second one was *spectral similarity* to reference drum sounds; that is to say, how similar are the phonemes' sounds are to those of real drums. Finally, the third one, and arguably the most important criterion, was *classification separability*, which evaluated how reliably algorithms can distinguish between different pairs of phonemes so as to maximise classification accuracy.

2 Related Work

To the best of our knowledge, this was the first study that addressed the problem of phoneme recommendation for online VPT, although earlier work touched on some aspects of our three criteria. Picart et al. [4] gathered data about the frequency of use of several phonemes among two beatboxers and Stowell et al. [5] released a dataset with fourteen beatbox performances from different participants with phoneme annotations in most of their sound events. In contrast with our case, these sound events are sometimes polyphonic (e.g. /ps/ = kick drum + opened hi-hat) and are heavily influenced by already established beatboxing techniques, although some of their insights were useful when approaching our own analysis.

Perhaps the most similar study to the present one regarding the spectral similarity between drum and phoneme sounds is that of Patel et al. [6], where the authors compared several audio features extracted from

	Kick Drum		Snare Drum		Closed Hi-Hat		Opened Hi-Hat		All Instruments	
/dʒ/	1	3.6%	-	-	-	-	-	-	1	3.6%
/k/	-	-	3	10.7%	-	-	2	7.1%	5	17.9%
/kʃ/	-	-	-	-	-	-	1	3.6%	1	3.6%
/kx/	-	-	3	10.7%	1	3.6%	-	-	4	14.3%
/p/	**22**	**78.6%**	3	10.7%	1	3.6%	-	-	22	78.6%
/s/	-	-	-	-	1	3.6%	2	7.1%	3	10.7%
/t/	4	14.3%	**7**	**25.0%**	**19**	**67.9%**	7	25.0%	**24**	**85.7%**
/ts/	-	-	3	10.7%	7	25.0%	**11**	**32.1%**	16	57.1%
/tʃ/	-	-	6	21.4%	4	14.3%	9	32.1%	14	50.0%
/tɕ/	-	-	4	14.3%	-	-	2	7.1%	6	21.4%
/tʒ/	1	3.6%	1	3.6%	-	-	-	-	2	7.1%
/ʔʕ/	1	3.6%	-	-	-	-	-	-	1	3.6%
/ǃ/	-	-	1	3.6%	-	-	-	-	1	3.6%

Table 1: Frequency of use of onset phonemes in the AVP dataset. These are the percentages and the raw numbers of participants (out of 28) that used the phonemes to trigger each of the four drum instruments (first four pairs of columns) and that used the phonemes to trigger at least one of the instruments ("All instruments" pair of columns).

both tabla sounds and their traditionally associated syllable sounds and found strong correlations between them, which suggests that onomatopoeia might have played an important role in the origin of such tabla vocables. Related recent studies focused on finding feature correlations between the acoustic space of drum sounds and that of their vocal imitations [7] and others tried drum sounds with their vocal imitations directly using both engineered features and features learnt by a neural network from the input spectrograms [8, 9]. Unlike the previous study in tabla sounds, neither of these carried out phoneme-wise feature analysis nor spectral analysis.

An especially relevant piece of research for classification separability and for this study in general was that of Stowell et al. [5]. In this work, which looked at online VPT with beatbox sound events, the authors explored classification accuracies under different frame delays from the onset times and also conducted a listening experiment on the perceived quality of different response delays so to provide an upper bound to the length of the analysis audio buffer. The accuracies they reported, however, took drum instruments as classes instead of phonemes, and therefore we could not extract phoneme-wise classification separability observations from such results. Another study [10] presented a VPT system with custom vocal percussion phonemes that could operate in online mode and also informed the users about the classification separability of their chosen sounds. This study did not include an investigation on phoneme-wise classification separability as well. However, a few studies in acoustic-phonetic analysis [11, 12] provided results on plosive phoneme classification separability. They used engineered features like the Mel Frequency Cepstral Coefficients (MFCC) and machine learning models like Hidden Markov Models (HMM) to classify plosive phonemes and illustrated results via confusion matrices. We got inspiration from such methods when planning the methodology of the present study; although, in contrast with our work, the whole phoneme sounds were taken for classification instead of just their first few milliseconds.

3 Analysis

3.1 Frequency of Use

To calculate the frequency of use of each phoneme in the AVP dataset, we first extracted the annotations regarding onset and coda phonemes of sound events contained in the personal subset. If we represent sound events as syllables, onset phonemes would be their first part, generally a plosive, fricative, or affricative phoneme. Coda phonemes would be the second part of the syllable, coming immediately after onset phonemes and usually being vowel phonemes. While coda phonemes are not necessary to build the sound

event, they come in handy for offline vocal percussion, providing a further degree of freedom to construct sound events upon. For the frequency of use criterion, we considered the onset phonemes of all sound events, with and without coda phonemes, while for the rest of the criteria we analysed those sound events that consist of just one onset phoneme without coda phoneme (see section 3.2).

We looked at how many participants used each onset phoneme to trigger each of the four instruments. For instance, how many participants used the onset phoneme /t/ to trigger the snare drum sound. The assumption we make here is that the higher the number of participants that decided to use a phoneme to trigger a particular drum sound, the more natural it would feel for a generic performer to trigger that drum sound with that phoneme sound. Results from this analysis are shown in Table 1.

There are two clarifications to be made regarding this table. The first one is that the "All Instruments" column contains the raw numbers and percentages of participants that used each phoneme. Note that this is not constructed simply by adding the terms in the rows of the four instruments, as one participant can use the same onset phoneme to trigger more than one drum instrument, in which case it would be still counted as one. The second clarification is that some participants ended up using two or more different onset phonemes interchangeably to trigger a single instrument, in which cases we included both phonemes. For instance, if we take a look at the kick drum column and sum up the numbers we get 29 participants (instead of 28). This is because the participant that used the phoneme /ʧ/ to trigger the kick drum also used the phoneme /ʤ/ for that purpose.

Looking at the table, we see that the dataset featured several plosive phonemes (/k/, /p/, /t/), fricative phonemes (/s/), affricative phonemes (/ʤ/, /kʃ/, /kx/, /ts/, /tʃ/, /tɕ/, /ʧ/, and /ʔʕ/), and even click phoenes (/!/). The phoneme /p/ was the preferred one to trigger the kick drum, /t/ to trigger the snare drum, /t/ again to trigger the closed hi-hat, and /ts/ to trigger the opened hi-hat. The phoneme /t/ was the one used by the highest number of participants (24) followed by /p/ (22), /ts/ (16), and /tʃ/ (14). It was also the only one used to trigger all four instruments. Interestingly, we found that the phonemes that participants used to trigger the snare drum were significantly varied compared to the rest of the instruments, which would deserve a closer look in future studies.

Here we also selected the most popular phonemes to be considered and analysed in the following sections. We carry out this filtering to both focus on the potentially most natural phonemes to do VPT with and to ensure representative sample sizes for the classification task. Onset phonemes that had a significantly similar method of vocal articulation were grouped together and reduced to one phoneme. This way, the phoneme /k/ accounted for the phonemes /k/ and /kx/; the phoneme /tʃ/ accounted for the phonemes /tʃ/ and /tɕ/; and /ʧ/ accounted for the phonemes /ʧ/ and /ʤ/. By defining such groupings, the "All Instruments" column in Table 1 changed accordingly and we got that the /k/ group was used by 8 participants, the group /tʃ/ was used by 18 participants, and the group /ʧ/ was used by 3 participants. In order to count the number of samples that each phoneme had, accounting for the newly-defined groups, we took the vocal percussion sound events associated with them that had no coda phonemes (i.e., pure onset phonemes). That way, the phoneme /k/ had 263 samples associated with it, the phoneme /p/ had 532, the phoneme /s/ had 23, the phoneme /t/ had 617, the phoneme /ts/ had 652, the phoneme /tʃ/ had 720, the phoneme /ʧ/ had 56, the phoneme /ʔʕ/ had 46, and the phoneme /!/ had 29. We observed that there was a pronounced jump between the number of samples associated with the phoneme /ʧ/ (56) to that associated with the phoneme /k/ (263). We considered the latter as an appropriate sample size for classification, with its associated phoneme (/k/) being used by more than a quarter of participants (28.6%). Therefore, we kept all phonemes whose sample size lied above 263 samples, leaving us with /p/, /k/, /t/, /ts/, and /tʃ/ as the final phoneme set to be analysed.

3.2 Spectral Similarity

This part of the process deals with the spectral analysis of the sound events relative to phonemes (/p/, /k/, /t/, /ts/, and /tʃ/) and reference drum sounds (kick drum, snare drum, closed hi-hat, opened hi-hat). For that, we take a single fixed-length frame from the first few milliseconds of each sound, compute its Fast Fourier Transform (FFT), and then apply an Equivalent Rectangular Bandwidth (ERB) scaling operation to the spectrum's frequency axis (ERB-rate scale) to derive the ERB spectrum [13]. This ERB scaling operation approximates the bandwidths of the filters in human hearing and, therefore, the kind of spectra derived from it would allow us to visually compare the sounds on the basis

of their perceptual similarity. Our hypothesis here is that the more similar the ERB spectrum of a certain phoneme is to that of a drum sound, the more natural it would feel to performers to trigger that specific drum by vocalising that phoneme.

We also had to select an appropriate frame size to compute the ERB spectrum from. We considered lengths of 512, 1024, 2048, 4096, and 8192 samples, which would be approximately equivalent to 12, 23, 46, 93, and 186 milliseconds respectively. We based our decision on two main properties of sound events. The first one is the sound events' effective duration, i.e., the amount of time that the sound's energy envelope lies above a certain threshold. The second one is how recognisable are sound events when cropped to different frame sizes, both visually and auditorily. For that, we took a look at the sounds' waveforms cropped at different lengths, listened to them, and discussed which length was the one that included the least amount of silence time while allowing samples to remain perceptually discernible. We concluded that 4096 samples was an ideal frame length to conduct the spectral analysis.

An important difference between the drum sounds in the BDF dataset and the phoneme sounds in the AVP dataset is that the former were recorded with professional microphones in a noise-isolated room, while the latter were recorded in an acoustically untreated room with a laptop microphone. This introduces noise when comparing the spectra of a drum sound to that of a phoneme sound. Fortunately, the AVP dataset includes a fifty-second audio file containing room noise; thus, in order to correct this issue to a reasonable extent, we subtracted the average magnitude spectrum of 500 non-overlapping noise frames to all the phoneme sounds' spectra, clipping results to zero so as to avoid negative spectral magnitudes. This effectively removed some of the components of the recordings' noise, including those belonging to the room and the laptop microphone.

After this, we took the average spectrum of the ERB spectra pertaining to a certain class (e.g. /p/ phoneme, snare drum,...) and we normalised it so that it had a minimum value of 0 and a maximum of 1. These ERB spectra, nine in total, are shown in Figure 1. We also compute the cosine similarities between individual ERB spectra from drum instruments and those from phoneme sound events. Table 2 collects the mean values of the resulting similarity matrices for each drum-phoneme combination.

	/p/	/k/	/t/	/ts/	/tsh/
Kick Drum	**.169**	.063	.042	.022	.033
Snare Drum	**.492**	.416	.303	.228	.303
Closed Hi-Hat	.312	.418	.512	**.522**	.437
Opened Hi-Hat	.321	.448	**.487**	.478	.440

Table 2: Mean cosine similarities between individual ERB spectra from drum instruments and those from phoneme sound events. Values in bold are the highest ones per drum instrument.

Several observations about sound similarity could be derived from Figure 1 and Table 2. For instance, the average spectrum taken from /p/ phonemes looks very similar to the one taken from snare drums. This means that it would potentially feel most natural for users to vocalise this particular phoneme to trigger a snare drum sound. The average spectrum of the /p/ phonemes also bears a marked resemblance to that pertaining to the kick drum, making it a second option to consider. Both these observations are corroborated by looking at the mean cosine similarities of the phoneme /p/ with the snare drum and the kick drum respectively.

The phoneme /k/ average spectrum, on the other hand, seems to resemble that of the opened hi-hat and, looking at the mean cosine similarities, it also appears to bear some similarity to the snare drum, achieving the second-highest similarity score for that particular instrument. We can also see that the average spectrum of phonemes /t/, /ts/, and /tʃ/ seem to be most similar to those of closed and opened hi-hat, especially to the former. This can also be corroborated by looking at the table with the mean cosine similarities, although by this metric the cosine similarity between the phoneme /tʃ/ and the opened hi-hat is slightly higher than the one between such phoneme and the closed hi-hat.

3.3 Classification Separability

An ideal online VPT system would trigger the response at the exact time that the vocal percussion sound event (the stimulus) begins; but, of course, this is impossible, as the algorithm has no prior information available to base its decision upon, and thus the need for an analysis buffer. Depending on the task at hand, one would require a longer or shorter maximum buffer length and, in the case of VPT, this maximum length is dictated by the perceptual unpleasantness of the system's response delay. Stowell et al. [5] conducted a

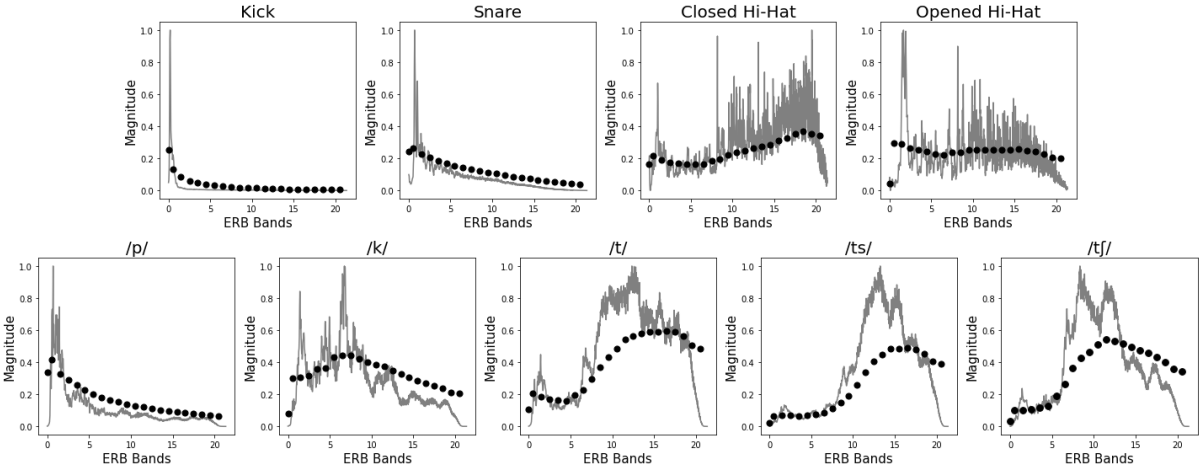

Fig. 1: Normalised mean ERB Spectra of the first 93 milliseconds of the four drum types and the five phonemes (grey) and mean value of each individual ERB band (black dots).

listening experiment in which a reactive system, right after the input is detected, outputs a mixture of all possible waveforms (kick, snare... etc) until a certain moment (e.g. 12 ms after) when this mixture waveform becomes the waveform of the correct class via crossfading. This showed that, for percussive events, listeners perceive delays as acceptable up to a buffer length of 35 milliseconds, which we adopted as a reference in our study. Therefore, a VPT system operating in online mode would have to detect and classify vocal percussion sound events within 35 milliseconds after they are produced by the performer.

Stowell et al. used 1024-samples long analysis frames with a hop size of 512 samples. These, at 44100 Hz of sample rate, are approximately equivalent to 23 and 12 milliseconds respectively. A delay of 35 milliseconds, as they refer to in their paper, would mean that the analysis frame of 1024 samples can be centered 1536 samples (35 milliseconds) from the onset time as most. If it is centered further than 1536 samples, performers would experience a delay in the system's response that is likely to not be well-tolerated perceptually speaking. Therefore, if a frame of 1024 samples is centered at 1536 samples after the onset, it means that we can analyse up to 2048 samples from the sounds' onset. Apart from this, we would also want to include some samples before the onset in our analysis buffer, as sound onsets are sometimes not accurately annotated or detected and some part of the sounds' transient might remain before

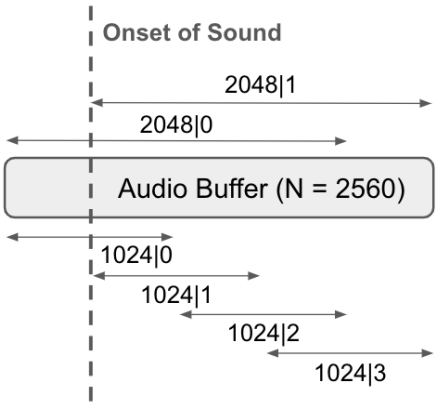

Fig. 2: Scheme of the extension and placement of the six frame types within the audio buffer.

the onset time. We take 512 samples (12 milliseconds) before the onset and incorporate them into the buffer. Hence, all in all, the length of the analysis buffer that we used to conduct online VPT was 2560 samples (47 milliseconds), which would be composed of the 512 samples immediately before the onsets and the 2048 samples after the onsets.

Our buffer could be analysed in several ways. For instance, we could take single analysis frames of different sample lengths centred at different points within the buffer and extract features from them, or even concate-

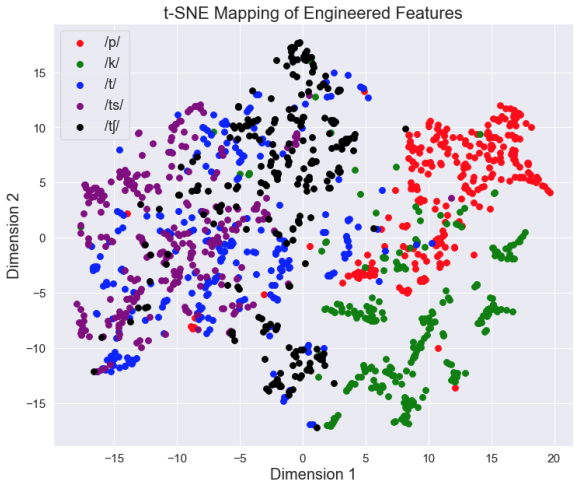

Fig. 3: Two-dimensional t-SNE mapping of phonemes' engineered features.

nate the features from different frames into a single feature vector. The latter approach was explored by Stowell et al. [5] and showed little to no improvement in classification separability. Therefore, taking into account that the higher the feature dimensionality the higher the algorithms' risk of overfitting, we decided to take single frames from the buffer and explore them individually. We took six frames whose positions and extensions in the buffer are illustrated in Figure 2 and explored their classification performance.

We extracted 38 engineered features from these six frame types. In the case of spectral features, we multiplied the frames by a Hann window before taking their FFT, from which we extracted the features. We extracted a total of 38 features, which included 13 Mel Frequency Cepstral Coefficients (MFCCs), spectral energy of 8 frequency bands (0-300, 300-800, 800-1600, 1600-4000, 4000-7000, 7000-11000, 11000-16000, and 16000-22050 Hz), 4 spectral roll-off frequencies (ratios of 0.25, 0.50, 0.90, and 0.95), spectral complexity, high frequency content, spectral strongpeak, 4 spectral central moments (centroid, variance, skewness, and kurtosis), spectral crest, spectral decrease, spectral entropy, spectral flatness, root mean square, and zero-crossing rate. We also reduced the sample size of the /p/, /t/, /ts/, and /tʃ/ phonemes to 263 via random sampling to ensure that all phoneme classes had an equal number of classes and thus have a balanced-class classification task.

We applied t-distributed stochastic neighbor embedding (t-SNE) [14] to reduce the dimensionality of the feature map from 38 to 2 so as to better visualise phoneme class separability. Figure 3 displays such feature map. There, we can see that the /p/ and /k/ phonemes are both moderately distinguishable from each other and very distinguishable from the /t/, /ts/, and /tʃ/ phonemes in a feature-wise sense. This was somewhat expected, as we could eventually see how their average spectra reflected this difference earlier in section 3.2. It is also worth noting how the data points pertaining to the phonemes /t/, /ts/, and /tʃ/ are very close together in the feature space, possibly meaning that they would be difficult for algorithms to separate. Although the phonemes /ts/ and /tʃ/ display moderate separability from each other, the data points relative to the phoneme /t/ are scattered all over the /ts/ and /tʃ/ regions, making this a possible conflicting phoneme when attempting classification.

The resulting feature vectors were normalised via the z-score before being fed to the classification algorithms. We explored 9 different machine learning-based algorithms (see Table 3) so as to provide an algorithm-independent measure of phoneme separability. We implemented the first eight algorithms via the Scikit-Learn library [15] and the ninth with the XGBoost library [16]. We performed hyperparameter optimisation by conducting a grid search for most of the algorithms' hyperparameters, applied 10-fold cross-validation to improve the generalisability and statistical significance of the output accuracies, and reported the best results from each method in Table 3. There, we could see which algorithm was potentially the best suited for separating the five phonemes and which of the six frame types allowed the algorithms to perform best in general. Most importantly, we wanted to visualise the classification separability of each phoneme with each other, which could be easily done via confusion matrices. We constructed two confusion matrices: one from the best-performing frame type's results averaged through the nine algorithms and another one from the best performance, both algorithmic and frame-wise. These two confusion matrices are displayed in Figure 4.

In Table 3, we see that the k-nearest neighbours algorithm performed best for all frame types but the 1024|0 one, where the extreme gradient boost performed best. Considering that the k-nearest neighbours algorithm simply operates by measuring the Euclidean distance between feature vectors, this highlights the general

	Fr. 2048|0	Fr. 2048|1	Fr. 1024|0	Fr. 1024|1	Fr. 1024|2	Fr. 1024|3
Nearest Centroid	**.673**	**.673**	.638	.662	.657	.670
Naive Bayes	.659	**.670**	.544	.649	.665	.624
Single-Layer Perceptron	.792	**.798**	.737	.778	.778	.784
Linear SVM	.782	**.792**	.736	.768	.773	.776
K-Nearest Neighbours	.830	**<u>.841</u>**	.751	.810	.817	.828
Decision Tree	.710	**.728**	.661	.700	.711	.714
Random Forest	.802	**.821**	.739	.789	.801	.809
Extreme Trees	**.704**	.693	.622	.692	.678	.653
Extreme Gradient Boost	.808	**.832**	.758	.803	.815	.804

Table 3: Best classification performances from each machine learning algorithm using each of the six frame types as inputs. Results are given in raw accuracy from 0 to 1. Bold numbers are the best performances with respect to the six frame types and the underlined bold number is the best performance overall.

a priori adequacy of the extracted features. Regarding frame types, we observe that the vast majority of best performances are achieved using the 2048|1 frame, i.e., the frame that extends from the onset time to the end of the audio buffer. This is somehow understandable, as this frame was the only one that covered the whole sound from its onset, so it had access to all the changes happening within that small portion of sound that maybe the rest of the frames could miss up to some point.

A more detailed view of the way the algorithms classified vocal percussion sound events into individual phonemes is given by Figure 4. We can immediately see that the observations that we drew earlier from Figure 3 perfectly translate to both confusion matrices. Essentially, the /p/ and /k/ phonemes are relatively separable from the rest, which lies in accordance to [12], while the phoneme /t/ is hard to separate from the /ts/, and /tʃ/ phonemes, especially in the case of the former.

4 Discussion and Conclusions

The results drawn from the previous analyses highlighted several properties of phoneme and drum sound events that were relevant for our goal, which was to choose the most appropriate phonemes to do online VPT with. In this section, we integrated the insights from these results and discussed the potential consequences that they may have in a real-life online VPT context. This ultimately led us to the final choice of phonemes which, as expected, depended significantly on the application setting.

The first part of the analysis, which looked at the *frequency of use* of onset phonemes in the AVP dataset, told us that the phonemes that people used the most were the /p/, /k/, /t/, /ts/, and /tʃ/ phonemes. This is the case for amateur vocal percussion, i.e., participants that have not learnt beatbox technique, but looking at the phonemes preferred by beatboxers we realise that they happen to be very similar to those chosen by amateur participants [5, 4]. This suggests that the choice of these phonemes to trigger certain drum samples is based on the vocal imitation of such drum samples to some extent and therefore the phonemes usually feel natural for the generic performer to use. We would therefore recommend using these phonemes for both online and offline VPT.

The second part of the analysis, which covered the *spectral similarity* among phonemes and drum sounds, gave us hints of which phoneme sound events typically sound more similar to a certain drum sound regarding their respective frequency distributions. We saw that the kick drum sound resembled the /p/ phoneme most, which explains the fact that the vast majority of participants (78.6%) selected such phoneme to trigger the kick drum sound. The average spectrum of the snare drum sound was also significantly similar to that of the /p/ phoneme, which is also reflected in their cosine similarity measure. This finding clashes with that of the frequency of use of the /p/ phoneme to trigger the snare drum (only 10.7% of participants) and the generally dispersed frequencies of use among phonemes for that particular drum sound. We suspect that this could be due to cultural conventions inspired by vocal percussion techniques like beatboxing or that it could also be a consequence of having already selected the /p/ sound for the kick drum and being inclined to choose

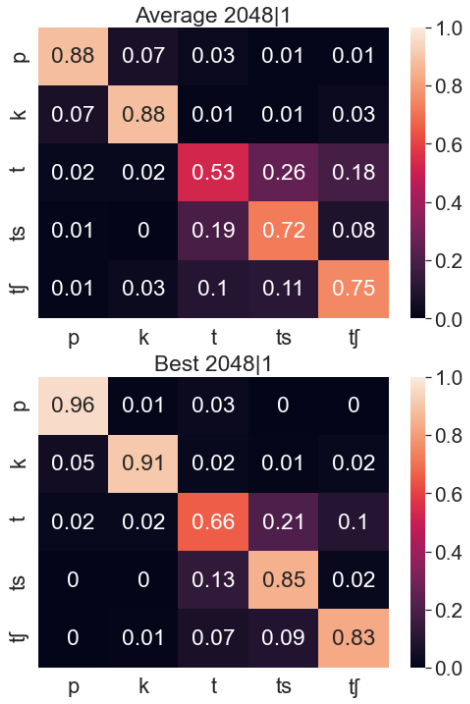

Fig. 4: Above: confusion matrix taken from the best-performing frame type (2048|1) averaged across all nine algorithms. Below: confusion matrix taken from the best classification performance (frame 2048|1 + k-nearest neighbours).

a different one for the snare drum. Either way, this would require a follow-up investigation that falls beyond the scope of this paper. Given that the /t/, /ts/, and /tʃ/ phonemes resembled closed and opened hi-hat sounds the most, we would be initially inclined to leave the /p/ sound to trigger the kick drum and the /k/ sound to trigger the snare drum, which were the first and the second most-featured phonemes for such instruments respectively.

The picture is completed via the third part of the analysis, which studied the *classification separability* among phonemes. Looking at the final confusion matrices, we saw that the phonemes /p/ and /k/ were generally well separated by the algorithms. This finding, along with the previous ones, provides a further reason to adopt these phonemes for the kick drum and snare drum respectively, as they are often used to trigger these instruments, they are spectrally similar to such instruments, and they can be separated in a classification process

K	p	**S, HO**	k, ts/tsh
S	p	**HC, HO**	ts, tsh
HC	t	**K, S, HC**	p, k, ts
HO	ts/tsh	**K, S, HO**	p, k, ts
K, S	p, k	**K, HC, HO**	p, ts, tsh
K, HC	p, t	**S, HC, HO**	k, ts, tsh
K, HO	p, ts/tsh	**K, S, HC, HO**	p, k, ts, tsh
S, HC	k, ts	**K, S, HC, HS, HO**	p, k, t, ts, tsh

Table 4: Phoneme recommendations for different drum set configurations. Notation: K = kick drum, S = snare drum, HC = closed hi-hat, HO = opened hi-hat, and HS = semi-opened hi-hat.

with relative ease. Also, the phoneme /t/ is especially hard to separate from the /ts/ phoneme and moderately hard to separate from the /tʃ/ phoneme. This tells us that, although the /t/ phoneme was the most featured one in the dataset, especially for hi-hat sounds, it might be best avoided so as to optimise classification performances.

In order to select the final phonemes for recommendation, we need to take into account the application context. In a VPT process, both offline and online, the users can select the number of instruments that they want to use for transcription. For instance, they could decide to only choose two sounds to trigger the kick drum and the snare drum, with no triggers for hi-hat sounds. Therefore, if the users could choose among four instruments and include any number of them in the set, we would have a total of 15 potential drum set configurations that would require individual analyses.

Also, the case of hi-hat sounds is usually a complex one. This instrument is able to make timbres that are very different from each other depending on how it is played. Here, we considered the fully closed and the fully opened case, while there also exists the half-opened sound of the hi-hat, which does not let the cymbals vibrate for as long as the fully opened hi-hat sound by bringing them closer to each other, where they clash and thus attenuate each other's sound.

Integrating the insights drawn from the previous discussion and taking the last paragraph's point into account, we built Table 4. This table lists all 15 possible drum set configurations along with their recommended trigger phonemes considering each case carefully, including a

16th combination featuring the semi-opened hi-hat as an extra instrument in case performers want to trigger it along with the other four drums. These recommendations are made considering the three criteria that we applied in the present paper, so we did not take into account the ease of their vocal articulation, for instance, or the skill level and consistency that a certain user may have when pronouncing these phonemes, which both fall beyond the scope of our paper. In any case, the phonemes featured in Table 4 are only meant to be recommendations for the users of the online VPT system, so it would be up to them to decide which phonemes to use in the end.

Future work will be focused on optimising the online classification performance of these phoneme sound events, gathering more vocal percussion sound events relative to the recommended phonemes and exploring data-driven methods with fast inference capabilities.

References

[1] Downie, J. S. and Hao, Y., "MIREX 2019 Evaluation Results," 2019.

[2] Delgado, A., McDonald, S., Xu, N., and Sandler, M., "A New Dataset for Amateur Vocal Percussion Analysis," in *Proceedings of the 14th International Audio Mostly Conference: A Journey in Sound*, pp. 17–23, 2019.

[3] InMusic, "BFD," `https://www.bfddrums.com/`, 2021, [Online].

[4] Picart, B., Brognaux, S., and Dupont, S., "Analysis and automatic recognition of human beatbox sounds: A comparative study," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4255–4259, IEEE, 2015.

[5] Stowell, D. and Plumbley, M. D., "Delayed decision-making in real-time beatbox percussion classification," *Journal of New Music Research*, 2010, doi:10.1080/09298215.2010.512979.

[6] Patel, A. D. and Iversen, J. R., "Acoustic and perceptual comparison of speech and drum sounds in the north indian tabla tradition: An empirical study of sound symbolism," in *Proceedings of the 15th international congress of phonetic sciences (ICPhS)*, pp. 925–928, 2003.

[7] Delgado, A., Saitis, C., Sandler, M., et al., "Spectral and Temporal Timbral Cues of Vocal Imitations of Drum Sounds," International Conference on Timbre, 2020.

[8] Mehrabi, A., Choi, K., Dixon, S., and Sandler, M., "Similarity measures for vocal-based drum sample retrieval using deep convolutional auto-encoders," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 356–360, IEEE, 2018.

[9] [Omitted], "Engineered vs Learnt Audio Features for Drum Sample Selection by Vocal Imitation," 2021, currently under review.

[10] Hipke, K., Toomim, M., Fiebrink, R., and Fogarty, J., "BeatBox: End-user interactive definition and training of recognizers for percussive vocalizations," in *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, pp. 121–124, 2014.

[11] Molho, L., "Automatic acoustic-phonetic analysis of fricatives and plosives," in *ICASSP'76. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pp. 182–185, IEEE, 1976.

[12] Sarma, B. D. and Prasanna, S. M., "Acoustic–phonetic analysis for speech recognition: A review," *IETE Technical Review*, 35(3), pp. 305–327, 2018.

[13] Glasberg, B. R. and Moore, B. C., "Derivation of auditory filter shapes from notched-noise data," *Hearing research*, 47(1-2), pp. 103–138, 1990.

[14] Van der Maaten, L. and Hinton, G., "Visualizing data using t-SNE." *Journal of machine learning research*, 9(11), 2008.

[15] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, 12, pp. 2825–2830, 2011.

[16] Chen, T. and Guestrin, C., "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.