# NONVERBAL SOUND DETECTION FOR DISORDERED SPEECH

*Colin Lea* *Zifang Huang* *Dhruv Jain*[*] *Lauren Tooley* *Zeinab Liaghat*
*Shrinath Thelapurath* *Leah Findlater* *Jeffrey P. Bigham*

Apple Inc.

## ABSTRACT

Voice assistants have become an essential tool for people with various disabilities because they enable complex phone- or tablet-based interactions without the need for fine-grained motor control, such as with touchscreens. However, these systems are not tuned for the unique characteristics of individuals with speech disorders, including many of those who have a motor-speech disorder, are deaf or hard of hearing, have a severe stutter, or are minimally verbal. We introduce an alternative voice-based input system which relies on sound event detection using fifteen nonverbal mouth sounds like "pop", "click", or "eh." This system was designed to work regardless of ones' speech abilities and allows full access to existing technology. In this paper, we describe the design of a dataset, model considerations for real-world deployment, and efforts towards model personalization. Our fully-supervised model achieves segment-level precision and recall of 88.6% and 88.4% on an internal dataset of 710 adults, while achieving 0.31 false positives per hour on aggressors such as speech. Five-shot personalization enables satisfactory performance in 84.5% of cases where the generic model fails.

***Index Terms***— Sound event detection, nonverbal communication, dysarthria, motor-speech disorders

## 1. INTRODUCTION

Many individuals with severe motor- or motor-speech disorders have limited communication ability and rely on ubiquitous technologies like phones and computers [1]. For people with motor impairments (e.g., carpal tunnel), assistive technology including voice control and eye tracking can be important parts of their daily life. Despite progress on disordered speech recognition [2, 3, 4, 5], commercial voice assistants have yet to be tuned for people with speech differences, so individuals with ALS, Muscular Dystrophy, Traumatic Brain Injury or other motor-speech disorders may rely on physical switch controls (e.g., buttons, sip & puff sensors, or joysticks) to interact with technology. These solutions can take orders of magnitude longer to accomplish the same tasks compared to people without motor disorders and may not be amenable to use in situations when an individual is laying in bed, outside of their wheelchair, or not at their desk [6].

We present a system for nonverbal, sound-based interactions that people with a wide range of speech disorders can use to interact with mobile technology. The input is raw audio and the output is a set of discrete events triggered when a user makes one of fifteen mouth sounds, such as "pop", "click", or the phoneme /i/, which can be used to perform actions like "select item" or "go back" on a mobile device. While conceptually simple, challenges arise when enabling robustness across wide vocalization ranges, achieving low-latency, and mitigating false positives from speech or background noises.

Prior work consists of early prototypes that are not robust to the needs of all-day consumer technology [7, 8, 9, 10, 11] or use sounds that do not suit all disabilities [12, 13]. Harada et al. [9, 10] developed an early system for people with motor-speech disorders which predicted vowels such as /u/ and /i/ and associated them with computer mouse motions. While valuable, speech or background noises (wheelchair sounds, music) could easily produce false positives. Recently, Cai et al. [12] introduced a system that is robust to the everyday needs of people with ALS, which solely detects the sound /a/, and is used to trigger actions like "call for help." They prevent false positives by also training with environmental sounds and by requiring a user to repeat the sound twice within ten seconds. While robust, the post-processor prevents real-time use cases. Talon Voice [13] and Parrot.py [11] are voice control libraries designed for tech savvy individuals with motor disabilities. Talon has two detectors ("pop" and "hiss"), which can trigger system events on a computer, but which are not sounds users with certain oral-motor function can vocalize. Parrot.py enables users to train custom detectors, but we find it is not robust to background sounds and can require tens or hundreds of training examples per detector.

We introduce a system that combines all of the benefits of the above work by: (1) using a universal sound set (i.e., all speaking individuals should be able to trigger at least one sounds, regardless of speech, accents, or other vocal characteristics), (2) providing robustness for all-day usage (i.e., not falsely triggering when someone is talking, music is playing, or loud environmental sounds are occurring), and (3) having low-latency (i.e., system interaction is on-par with touch-based systems). We describe a dataset, model, and a training

[*] University of Washington (work done during an internship at Apple)

| Nonspeech | Definition |
|---|---|
| Click | Tongue to roof of mouth (front), snap down. |
| Cluck | Tongue to roof of mouth (back), snap down. |
| Pop | Close lips tightly and release with quick blow. |
| **Voiceless** | **Definition** |
| /p/: Pitch | Close lips loosely and blow out. |
| /k/: Kite | Touch back of tongue to roof of mouth, exhale. |
| /t/: Teeth | Open lips, teeth closed. |
| /ʃ/: Shoe (“sh”) | Push lips out with your teeth open and blow air. |
| /s/: Snake | Open lips with your teeth closed and blow air. |
| **Voiced** | **Definition** |
| /ɛ/: Effort (“eh”) | Open mouth, tongue raised, and start voicing. |
| /ə/: Ump (“uh”) | Open mouth, teeth open, tongue slightly raised. |
| /u/: Boo (“oo”) | Form an O with your lips. |
| /m/: Mom | Start voicing with lips closed. |
| /i/: Eagle (“ee”) | Open mouth with lips wide, tongue slightly raised, and start voicing. |
| **Diphones** | **Definition** |
| /la/: Law | Touch tongue to roof of mouth. Start voicing and open mouth. |
| /mə/:Mud(“muh”) | Close lips, start voicing, and open your mouth. |

**Table 1**: Nonverbal Sound List. IPA and English forms are used interchangeable in text for reader convenience.

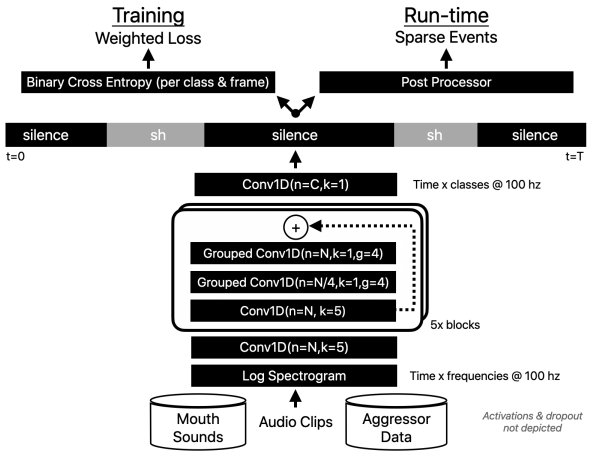


**Fig. 1**: During training, predict the probability of each sound per-frame, using mouth sounds and aggressor audio (speech, environmental sounds). At test time, take these probabilities and generate sparse events. $k$=width, $g$=groups, $n$=nodes.

scheme to improve robustness across variations in vocalizations, – including via personalization – and evaluate on data from individuals with and without motor-speech disorders.

## 2. NONVERBAL SOUNDS

### 2.1. Sound Types & Clinical Relevance

With clinical guidance, we looked at prototypical examples of speech production for individuals with cerebral palsy, ALS, musclar dystrophy, multiple sclerosis, traumatic brain injury, and other conditions resulting in speech disorders. We identified 15 sounds spanning non-speech, voiced phones, unvoiced phones, and diphones as shown in Table 1.

Sounds were chosen based on features (i.e., voicing, nasality/resonance) that people with specific diagnoses are more likely able to produce while maintaining diverse locations of production in the oral cavity (palatal, alveolar, bilabial, velar) to ensure success for a large distribution of people. Vowels “eh”, “ee” and “oo” were chosen for their spectral differences and because some (i.e., “ee” and “eh”) may be more intelligible in individuals with ALS than other vowel choices [14]. The central vowel “uh” may be more easily produced for individuals with cerebral palsy and others with dysarthria [15]. “Muh” was chosen as a consonant-vowel (CV) production that is easier for individuals who tense their oral structures when initiating speech, while maintaining the central vowel “uh”. /m/ in isolation and in “muh” may be more clearly produced for people who may have hypernasality (i.e. due to flaccid dysarthria or when wearing BiPap for respiration). Individuals who are unable to phonate consistently (e.g., due to respiratory incoordination, ventilator dependence, or a voice disorder) may be able to produce non-voiced phones such as /k/, /p/, /t/, /s/, “sh”, or non-speech sounds “click”, “cluck”, and “pop”.

### 2.2. Datasets

Despite the short duration of our sounds, there is large variation in pronunciation across accents, ages, genders, and vocal abilities. In contrast with (e.g., [12]) which train and evaluate isolated vowel detectors on public English speech datasets, we collected over 100k instances of isolated sounds using the protocol described below. We also use a large and diverse set of aggressor data, in the form of speech and environmental sounds, to prevent false detections in everyday situations.

Mouth Sounds: We collected audio from 710 non-disabled people with at least 40 participants each across demographics spanning accent/locale, age ranges (18+), gender (male, female, non-binary), device type (phone, tablet, wired or bluetooth headphones), and background environment (indoor or outdoor). Each person recorded audio clips of themselves repeating each sound type at least 10 times in a row with about one second of silence between vocalizations. Recordings were done at a “close” and “far” distance to simulate holding a device in hand and speaking into a tablet potentially mounted on a wheelchair or table.

Obtaining data across accents and physical locations is important. Early models trained on predominantly US accents achieved 24.7% worse F1 score compared to the same models trained on people with nine accents (British, Chinese, French, German, Indian, Italian, Japanese, Spanish, US). One mode of variation was from people whose native language (e.g., Italian) only had five scripted vowels instead of the seven

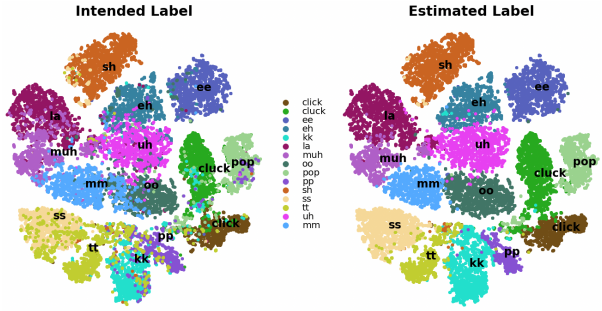


**Fig. 2**: T-SNE visualization of sound event embeddings using the model in Section 3. Embeddings are colored using (Left) the sound type assigned by a participant and (Right) the sound type assigned by our model. This is used to identify discrepancies in how people vocalized each sound and the label type.

used in English (i.e., "uh" and "eh" are used interchangeably). We also found larger variation in how people from different countries tended to say each sound, regardless of our written and visual descriptions (i.e., "uh" was sometimes pronounced "oo"). These discrepancies were apparent when listening to clips and when visualizing similarity of their sound embeddings, as shown via the T-SNE plot [16] in Figure 2. We automatically detected discrepancies using two rounds of Pseudo Labeling [17] and found 9.8% of self-described labels to be different than our prediction including 11.3% from "eh" to "uh", 8.5% "uh" to "oo", and 6.5% "ee" to "eh". These clips were removed when training final models.

Aggressors: We train and evaluate using speech and non-speech datasets to mitigate false positives. For speech, we use subsets of LibriSpeech which contains read speech [18] (*train-clean-100* and *test-clean*), public podcast recordings of people with US and British accents, and 10 phrases from each participant in our mouth sounds collection. For non-speech data we rely on environment sounds from AudioSet [19] and internally collected recordings such as appliance sounds. Training clips are randomly sampled from each dataset, totaling 30 hours of speech and 20 hours of background sounds.

Annotations: Each mouth sound recording contains repeated instances of one sound type with silence in between. Frame-wise labels were generated by computing the energy in the audio signal and finding segments with minimum duration of 30 ms and whose relative energy exceeded one standard deviation from the mean. All frames within a given segment were labeled with the user-annotated sound type and all others were considered "silence." Labels for speech clips were generated using a speech activity detector and all aggressor clip frames were labeled with the background class.

## 3. LOW-LATENCY SOUND DETECTOR

Our system is visualized in Figure 1. A preprocessor computes log spectrograms, a temporal convolutional network computes the probability that each frame contains a sound, and a post-processor takes probabilities and outputs sparse detections. At run time all modules are applied at 100 hz.

Model Architecture: Our model is a simple Temporal Convolutional Network, most similar to QuartzNet [20]. The input 64 dimensional log mel-spectrograms generated from 16k hz audio with a 25 ms window and stride of 10 ms, resulting in a 100 hz sampling rate. The first layers apply 1D convolutions (kernel size `k=5`) with `N=256` nodes. There are then five blocks of grouped (`g`) convolutions with the following pattern: `Conv1D(n=N,k=5,g=4)`, `LeakyReLU`, and a residual bottleneck consisting of `Conv1D(n=N/4,k=1)`, `LeakyReLU`, `Conv1D(n=N,k=1)`. Dropout is used after each activation. The network head consists of a `Conv1D(n=C,k=1)` with `Sigmoid` activation. Each frame's output is a vector of size $C = 17$: 15 nonverbal sounds, a background class, and a speech class. The receptive field is 270 ms.

Post-processing: Many of our sounds are similar to what appear in everyday speech. We prevent false positives using a post-processor that aggregates background, speech, and nonverbal probabilities and outputs sparse events $(c, t)$ for class $c$ and time $t$. For each class, given probabilities $p_{c,t}$ for times $1...t$, generate an event if $p_c$ is greater than threshold $\theta_c$ for the most recent $\tau_c$ frames. No event is generated if the background or speech probabilities exceed $\theta_{bg}$ in the past 50 frames or if any class is detected within this time.

Post-processing parameters are optimized per-class to minimize the weighted F1 score on mouth sounds data, False Positive Rate on speech aggressor data, and latency. Optimal values range from $\theta_c \in [0.4, 0.6]$ and $\tau_c \in [7, 15]$ frames. Sounds including click and pop may be 50 ms whereas /u/ or /i/ may be 250 ms, and values of $\tau_c$ reflect this. If additional robustness is required, additional processing can be used to further reduce false positives by requiring silence after each sound, albeit at the cost of added latency.

### 3.1. Model Training

Baseline models are trained using a binary cross entropy loss per-class. Batches of 50% mouth sound clips and 50% aggressors are concatenated, with cumulative duration of $T$ frames, outputting $T$ log probability vectors, with a loss evaluated at 100 hz before the post-processing function. Boundaries of each segment are inflated by 50% of the receptive field size (13 frames) to encourage the model to detect the onset and offset of a sound, where many of the constituent frames are "silence". This is equivalent to the temporal augmentation used by Meyer et al. [21].

Personalization: Our datasets contain predominantly non-

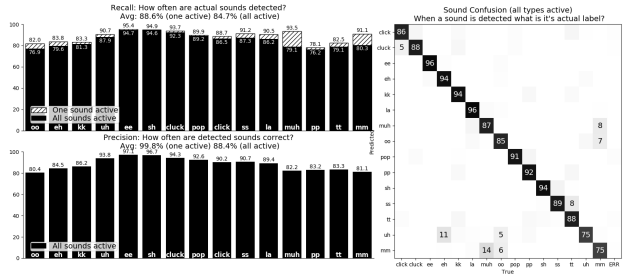


**Fig. 3**: Segmental precision/recall on our 90 person non-disabled set. "One active" means only one sound type is enabled at a time. 'All active" means any class can be detected.

disordered speech, and there is risk that the system does not work as well for users with severe speech differences. We investigated whether models could be personalized by fine-tuning on example vocalizations from a user. We use 256-dim embeddings from the pre-trained model above and fine tune on recordings of someone repeating the same sound one to five times. Weights in the final class-specific layer are updated using the automated labeling scheme described above and using a frame-wise binary cross entropy loss. Models are trained using vocalizations from one recording and evaluated using a separate clip typically recorded 15 minutes later.

## 4. EXPERIMENTS & ANALYSIS

Baseline Results: Figure 3 (top) shows segmental precision and recall metrics and a confusion matrix on a 90 person mouth sound evaluation set (5 male & 5 female per accent). A segment is considered correct if the model detects the correct event anywhere between the start and end of a sound. In practice, someone using this type of feature may only use a few sound types per session; they likely will not need all 15 detectors at the same time. As such, results are shown for the extremes where only one detector is active ("one active") and or all detectors ("all active"). The biggest discrepancy is for "mm" and "muh", which are often confused if both are active, but achieve high performance when used individually.

Personalization: Experiments were performed on participants for whom the generic model fails (i.e., $F1 < 50\%$). Fine-tuning on one, three, or five examples from that user improves F1 by 55.8%, 58.9%, and 61.8% on held out recordings. 84.5% of sounds that could not be detected with the generic model could be detected after personalization with five samples. "click" (74.3%), "pop" (68.5%), and "oo" (68.4%) have the largest improvements. Investigations with MAML and ProtoNets using [22] did not yield significant improvements.

Aggressors: Our final model, trained using "positive" mouth sounds and "negative" speech/background sounds, has 4.65 false positives (FPs) per hour on LibriSpeech test-clean. "Sh" and "uh" have higher rates (0.56 & 0.74 FP/hour) whereas click and "mm" have zero. We trained the same model without these negative datasets and it has 303.9 FPs/hour. Thus, this simple technique reduced the false positive rate by 98.4% while losing only 0.6% and 1.8% precision and recall on mouth sounds data. A similar experiment on a 10.5 hour environmental sound set (e.g., kitchen noises) reduced the false positive rate from 238.5 to 0.225 FP/hour. Speakers and environments did not overlap in the training and test sets.

Latency: The average system latency is 108±32 ms from the end of each vocalization to system detection. Extending the boundaries of each label as described in Section 3 improves sound onset detection and reduces latency by 33 ms. Our model starts to detect sounds before they have been fully vocalized, which means that longer sounds such as /s/ or "sh" are sometimes detected before completion. The computation time on an iPhone 12 is approximately 1 ms so the total amortized latency is within the range of typical touchscreen interactions (50-200 milliseconds [23]).

Motor-Speech Results: Recordings and feedback were collected from 28 people with speech differences resulting from cerebral palsy, muscular dystrophy, dysphonia, Parkinson's disease, or another motor-speech disorder. Four had moderate-to-severe speech disorders as judged by speech intelligibility and the remaining had mild. Individuals tested a real-time version of this work and recorded themselves making each sound 10 times for quantitative evaluation. The average success rate (F1≥50%) was 82%. Lowest performing sounds were $/k/$ (68%), $/s/$ (72%), $/t/$ (72%). For 23 people, at least 10 of 15 sounds were successfully detected. Errors sometimes resulted when a sound was vocalized slowly relative to the non-disabled population or when an individual needed to vocalized sounds like $/t/$ as "t-uh" sometimes due to their speech difference. For an individual with very low speech intelligibility only 3 of 15 sounds could be detected. For a user on a breathing apparatus, some sounds (e.g., "sh") did not work while the apparatus was active. This issue was mitigated by using other sounds that were not impacted such as $/k/$. Individuals with Parkinson's disease indicated that they would be interested in this feature during the times of the day when symptoms are most severe. Some individuals report needing to be close in proximity to the device due to a limited ability to vocalize loudly. We received positive feedback from people who have used our work in situations where they otherwise cannot interact with technology for mobility reasons (e.g., in bed or when not in their wheelchair).

## 5. CONCLUSION

We developed a system for nonverbal sound detection using triggers like pop and click that is robust to everyday interactions with background speech and environmental noise. This was designed to work for a wide range of vocal abilities and was validated on people with and without speech disorders.

## 6. REFERENCES


[1] S Koch Fager, Fried-Oken M, T Jakobs, and DR Beukelman, "New and emerging access technologies for adults with complex communication needs and severe motor impairments: State of the science," in *Augmentative and Alternative Communication*, 2019.

[2] JR Green, B MacDonald, PP Jiang, J Cattiau, R Heywood, R Cave, K Seaver, M Ladewig, J Tobin, M Brenner, PQ Nelson, and K Tomanek, "Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases," in *Interspeech*. ISCA, 2021.

[3] J Harvill, D Issa, M Hasegawa-Johnson, and C Yoo, "Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary," in *ICASSP*. IEEE, 2021.

[4] M Kim, Y Kim, J Yoo, J Wang, and H Kim, "Regularized speaker adaptation of kl-hmm for dysarthric speech recognition," *IEEE*, 2017.

[5] F Rudzicz, "Articulatory knowledge in the recognition of dysarthric speech," *IEEE*, 2011.

[6] S Kane, A Guo, and MR Morris, "Sense and accessibility: Understanding people with physical disabilities' experiences with sensing systems," in *ACM ASSETS*, October 2020.

[7] T Igarashi and JF Hughes, "Voice as sound: using non-verbal voice input for interactive control," in *UIST*, 2001.

[8] M Funk, V Tobisch, and A Embfield, "Non-verbal auditory input for controlling binary, discrete, and continuous input in automotive user interfaces," in *CHI*, 2020.

[9] S Harada, JA Landay, J Malkin, X Li, and JA Bilmes, "The vocal joystick: evaluation of voice-based cursor control techniques," in *ACM SIGACCESS*, 2006.

[10] S Harada, JO Wobbrock, and JA Landay, "Voicedraw: a hands-free voice-driven drawing application for people with motor impairments," in *ACM SIGACCESS*, 2007.

[11] "Parrot.py," `https://github.com/chaosparrot/parrot.py`, Accessed: 2021-09-26.

[12] S Cai, L Lillianfeld, K Seaver, JR Green, MP Brenner, PC Nelson, and D Sculley, "A voice-activated switch for persons with motor and speech impairments: Isolated-vowel spotting using neural networks," *Interspeech*, 2021.

[13] "Talon voice," `https://talonvoice.com`, Accessed: 2021-09-26.

[14] J Lee, E Dickey, and Z Simmons, "Vowel-specific intelligibility and acoustic patterns in individuals with dysarthria secondary to amyotrophic lateral sclerosis," in *J of Speech, Language, and Hearing Research*, 2019.

[15] BM Ansel and RD Kent, "Acoustic-phonetic contrasts and intelligibility in the dysarthria associated with mixed cerebral palsy," in *J of Speech, Language, and Hearing Research*, 1992.

[16] L Van der Maaten and G Hinton, "Visualizing data using t-sne," *J of Machine Learning Research*, 2008.

[17] DH Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *ICML Workshop*.

[18] V Panayotov, G Chen, D Povey, and S Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015.

[19] JF Gemmeke, DPW Ellis, D Freedman, A Jansen, W Lawrence, RC Moore, M Plakal, and M Ritter, "Audioset: An ontology and human-labeled dataset for audio events," in *ICASSP*. IEEE, 2017.

[20] S Kriman, S Beliaev, B Ginsburg, J Huang, O Kuchaiev, V Lavrukhin, R Leary, J Li, and U Zhang, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *ICASSP*, 2020.

[21] J Meyer P Warden VJ Reddi M Mazumder, C Banbury1, "Few-shot keyword spotting in any language," in *Interspeech*, 2021.

[22] A Arnold, P Mahajan, D Datta, I Bunner, and KS Zarkias, "learn2learn: A library for Meta-Learning research," Aug. 2020.

[23] A Ng, J Lepinski, D Wigdor, S Sanders, and P Dietz, "Designing for low-latency direct-touch input," in *User Interface Software and Technology*, 2012.