HHS Public Access

Author manuscript

IEEE EMBS Int Conf Biomed Health Inform. Author manuscript; available in PMC 2020 August 27.

Published in final edited form as:

IEEE EMBS Int Conf Biomed Health Inform. 2019 ; 2019: 1–4. doi:10.1109/bhi.2019.8834506.

DeepDDK: A Deep Learning based Oral-Diadochokinesis Analysis Software

Yang Yang Wang[*], **Ke Gao**[*], **Yunxin Zhao**[*], **Mili Kuruvilla-Dugdale**[†], **Teresa E. Lever**[‡,1], **Filiz Bunyak**[*,1]

[*]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri 65211

[†]Department of Speech, Language and Hearing Sciences, University of Missouri, Columbia, Missouri 65211

[‡]Department of Otolaryngology - Head and Neck Surgery, University of Missouri, Columbia, Missouri 65211

Abstract

Oromotor dysfunction caused by neurological disorders can result in significant speech and swallowing impairments. Current diagnostic methods to assess oromotor function are subjective and rely on perceptual judgments by clinicians. In particular, the widely used oral-diadochokinesis (oral-DDK) test, which requires rapid, alternate repetitions of speech-based syllables, is conducted and interpreted differently among clinicians. It is therefore prone to inaccuracy, which results in poor test reliability and poor clinical application. In this paper, we present a deep learning based software to extract quantitative data from the oral DDK signal, thereby transforming it into an objective diagnostic and treatment monitoring tool. The proposed software consists of two main modules: a fully automated syllable detection module and an interactive visualization and editing module that allows inspection and correction of automated syllable units. The DeepDDK software was evaluated on speech files corresponding to 9 different DDK syllables (e.g., “Pa”, “Ta”, “Ka”). The experimental results show robustness of both syllable detection and localization across different types of DDK speech tasks.

Keywords

Diadochokinesis analysis; speech signal analysis; deep learning; event detection; event localization

I. Introduction

Diagnostic and prognostic accuracy as well as timely intervention and treatment monitoring are important for progressive neurological disorders such as Parkinsons disease (PD), amyotrophic lateral sclerosis (ALS), and multiple sclerosis (MS), since earlier intervention is associated with improved quality of life and survival in these patient populations. Diagnosis and monitoring of neurological disorders involve various medical tests, some of

[1]Corresponding author. levert@health.missouri.edu, bunyak@missouri.edu.

which can be invasive and expensive, prohibiting their effective use. Recent advances in mobile health technologies have lead to the development of non-invasive, more accessible, and affordable new methods and devices not only for diagnosing and monitoring medical conditions, but also for tracking functional decline induced by these diseases. This paper focuses on development of an oral-diadochokinesis (oral-DDK) analysis software for non-invasive, objective, and quantitative assessment and monitoring of speech disorders that are common in PD, ALS, MS, and other neurological disorders.

Oral-DDK tasks are universally used by speech-language pathologists (SLPs) for assessment and monitoring of motor speech disorders (e.g., dysarthria and apraxia) [13]. These tasks involve repetitions of single syllables like “Pa”, “Ta”, “Ka”, or sequential multi-syllables such as “Pa-Ta-Ka ”, “Buttercup”, etc. as fast as possible, in one breath or within a fixed period of time. SLP use these tasks to estimate diadochokinetic (DDK) rate to provide information about a person’s ability to make rapid speech movements using different parts of the mouth [5]. Manual analysis of DDK rate from audio files is subjective, time intensive, and error-prone. Furthermore, since manual analysis only estimates syllable count, not the locations (timestamps), or production accuracies of the events, rich information that can help diagnosis or monitoring is lost.

In this paper, we present an oral-DDK analysis software that can detect syllables and estimate their timestamps in DDK audio files. The software consists of two main modules, a deep learning based fully automated syllable detection module, and an interactive visualization and editing module that allows inspection and correction of automated syllable detection results. The aim of this work is to enable computation of objective, quantitative outcome measures from the oral-DDK signals to aid early diagnosis and treatment monitoring of neurological disorders.

II. Related Work

Recent studies have started to show the potential of speech in general and oral-DDK in particular to be a functional biomarker for neurological disorders. DDK task derived measures were explored for the diagnosis of PD [4] [7], traumatic brain injury [3], MS [2], ataxic dysarthria [1], etc. using a variety of computational approaches.

For example, syllables in oral-DDK task can be detected by first computing the signal envelope, then by thresholding the envelope or locating local maxima in the envelope. This process requires parameter selection for envelope computation and thresholding. However, complexity of the signals, and high variations in frequency and amplitude (Figure 1) make parameter selection challenging and result in under or over-detection of the syllables. Wang et al. [11] proposed a multi-threshold syllable detection system in which a threshold is automatically selected based on a 7-second DDK sample and the gender of the participant. Threshold can be adjusted to re-perform the analysis if needed. However, if the lowest peak intensity during consonant-vowel (CV) is lower than the highest peak intensity during inter-syllable pauses, the DDK sample gets labeled as nonexecutable. The approach results in more than one third of their DDK samples being unanalyzable. Tao et al. [8] proposed use of Gaussian Mixture Models and Hidden Markov Models (GMM-HMM) to automatically

detect syllable boundaries in DDK data. In recent years, deep learning based audio analysis has started to become popular, such as [6] [12] [9]. In this paper, we propose a deep learning based audio signal analysis system for automated detection and localization of syllables in oral DDK samples.

III. METHOD

Deep learning is a subfield of machine learning that allows learning of high-level abstractions in data through its multilayer architecture [10]. Inspired by the recent successes of deep learning in speech and image analysis, we have developed DeepDDK, a deep learning based system for automated detection and localization of syllables in oral-DDK tasks.

A. Network Architectures

DeepDDK consists of a cascade of two convolutional neural networks (CNNs). The first CNN (CNN-1) segments the 1D audio signal into syllable vs. non-syllable (silence, breath, etc.) regions. CNN-1 is a classification network that operates on 1D temporal array of audio samples. Input size is 1 × 5292, where $5292 = 120ms \times 44.1kHz$ corresponds to the product of average syllable duration and sampling rate. Output size is two, corresponding to syllable and non-syllable labels. The CNN-1 network structure is shown in Figure 2a and Table I.

The second CNN (CNN-2) locates syllable timestamps within the syllable regions detected by CNN-1. CNN-2 is a 2D regression network that operates on a sequence of audio frames (temporal windows) with length of average event duration $120ms$ in Table II to predict the precise timestamp of a syllable. For each frame in the input, CNN-2 predicts the probability to contain a syllable. Input size is 15×5292, where 15 is the number of audio frames analyzed and 5292 is the length of a frame as in CNN-1. The CNN-2 network structure is shown in Figure 2b and Table I.

B. Convolutional Neural Networks Training

Using our custom DDK data collection iOS App, we conducted an IRB-approved study to collect oral-DDK data from seventeen testers for nine tasks (corresponding to syllables “Pa”, “Ta”, “Ka”, “Da”, “Ba”, “Ga”, “La”, “Ma”, and “Ha”). Following study consent, subjects were instructed to repeat each syllable as fast as they could for 15 seconds. Each task was repeated twice, resulting in 306 audio files of length 15 seconds. All audio files were sampled at 44.1 kHz. Our DeepDDK system relies on availability of labeled training data. In order to label data, we have developed a preliminary unsupervised automated syllable detector (envelope with local maxima) with a user interface for visualization, navigation, and editing of the results. Results form unsupervised detector were inspected by three experts and corrected according to consensus using our visualization and editing interface. The ground truth consists of a timestamp for each syllable, instead of a region in the audio signal. Locations of these timestamps typically correspond to the sample value maxima in the syllable/event. Instead of using a sliding window, considering average event duration (see Table II), a 120ms temporal window (also called ‘frame’) centered around each syllable timestamp is used as positive sample (syllable) to train CNN-1 as well as to reduce false

positive training data. So 70 timestampes will have 70 positive frames. Shaded regions in Figure 3 show these positive samples. As negative samples (non-syllable) to train CNN-1 and to reduce false negative training data, 120ms temporal windows centered at each midpoint between two consecutive syllable timestamps are used. In order to prevent information loss, instead of extracting hand-crafted audio features, raw audio sample value is fed into CNN-1. CNN-2, the regression module, aims to predict precise timestamps of the detected syllables. To train the network, 15 sequential frames (one centered on, K centered before, and $15 - K - 1$ centered after the ground truth timestamp, where K is a random number in the range 1 to 13 for robustness) are extracted with a step size of 12ms. Each frame is assigned a score indicating its probability to contain a syllable:

$$P_i = 1 - \frac{|i - i_{GT}|}{15 - 1} \quad (1)$$

where i is the index of the specific frame in the sequence and i_{GT} is the index of the frame centered on the ground-truth syllable timestamp.

C. Convolutional Neural Networks Testing

DeepDDK syllable detection and localization processes can be summarized as follows. The intermediate outputs from classification network (CNN-1) and regression network (CNN-2) are shown in Figure 4b and 4c.

Step-1 Classification: Raw audio sample value is fed into CNN-1. For each sliding window with stride 12ms, CNN-1 predicts a class label (syllable vs. non-syllable), which is then assigned to the sliding window. The process produces a binary 1D array, $\mathscr{L}$, where $\mathscr{L}(t) = 1$ indicates presence of a syllable at time t.

Step-2 Interval Preprocessing: Morphological closing is applied to $\mathscr{L}$ to fill small gaps in class labels. Syllable event time intervals $\mathscr{E} = \{\mathscr{E}_1, .., \mathscr{E}_n\}$ are identified by applying connected component labeling to $\mathscr{L}$. $\mathscr{E}_i$ represents a region of a syllable. n indicates syllable/event count in the file.

Step-3 Syllable Timestamp Prediction: From each syllable event interval $\mathscr{E}_i$, 15 sequential frames are extracted. If the duration of $\mathscr{E}_i$ is less than 15 frames, the negative frames around $\mathscr{E}_i$ will be included until $\mathscr{E}_i$ duration has 15 frames. Extracted frames are fed into CNN-2 for timestamp score prediction. The center of the frame with the maximum CNN-2 score is marked as the timestamp for event $\mathscr{E}_i$.

IV. Experimental Results

As described in Section III-B, we have collected 306 audio files corresponding to seventeen subjects, nine different syllable/event types, and two files for each event type. These files were first analyzed by our unsupervised gammatone-based syllable detection program. The detections were then corrected by expert speech pathologists using our visualization and editing interface to produce ground-truth data. Out of these 306 files, 225 files (74%) were

used to train the proposed DeepDDK software, and 81 files (26%) were used to test the syllable detection and localization performance. Each audio file was 15sec long. The average number of events per audio file was 74.

We evaluated the system performance in terms of syllable/event count accuracy and syllable/event localization accuracy. Event counting accuracy is evaluated by comparing the number of detected events (DT) to the number of ground truth events (GT). The average event count difference $\frac{1}{N}\sum_{i=1}^{N}|\#\,DT(i) - \#\,GT(i)|$ between DeepDDK and ground-truth for $N = 81$ test files is 1.6 events. The average execution time per test file is 1.9s. Figure 5 shows detailed, comparative, syllable count accuracy analysis for the proposed DeepDDK and a very recent pre-linguistic speech segmentation tool [14]. DeepDDK results in low syllable count errors and outperforms the pre-linguistic speech segmentation tool [14]. For 81% of the test files, DeepDDK count error is 2 or less ($|\#DT(i) - \#GT(i)| \leq 2$). Considering that the average number of events per file is 74, this corresponds to 2.70% error. For the case of [14], only 72% of the test files have a count error of 2 or less. Figure 5 also shows that DeepDDK's highest error for any file is 5, which corresponds to an error of 6.75%, whereas when [14] is used, 17% of the files have a count error higher than 5. We also compared our results with another DDK software from Smekal et al. [15] [16], and linear support vector machine (SVM) with Mel-frequency cepstral coefficients (MFCC) features. Table II presents overall and type-specific event localization performances for Deep-DDK. Localization performance is measured in terms of recall ($Recall = \frac{\#\,TP}{\#\,GT}$) and precision ($Precision = \frac{\#\,TP}{\#\,DT}$) for a given temporal distance threshold. Two temporal distance thresholds $T = 30ms$ and $T = 60ms$ were used to evaluate performance. If the timestamp of the detected event is located within T milliseconds of the ground-truth event, the detection is considered a true positive detection (TP). We can see in Table II that event types 'Pa', 'Ka', 'Ba', 'Da', and 'Ga' have very high location accuracies for $T = 30ms$, because of their fairly regular pattern. In contrast, 'Ta' appears to have lower location accuracy. However, this is mostly due to its relatively longer duration (larger than our frame length), which leads to the shift location of the predicted event label.

V. CONCLUSIONS

We have presented DeepDDK, a deep learning based system for automated analysis of oral-DDK tasks. DeepDDK allows objective and quantitative analysis of oral-DDK data, corresponding to a task used by SLP for assessment and monitoring of abilities. Experimental results showed robust syllable detection and localization capabilities across different types of DDK. Accurate, objective, quantitative analysis of oral-DDK data is of great significance because these tasks can be potentially used in diagnosis and monitoring of disorders, particularly the progressive ones such as PD, ALS, and MS.

Acknowledgments

This work was funded by the University of Missouri Coulter Translational Partnership Program.

References

[1]. Horne Malcolm, Power Laura, and Szmulewicz David. “Quantitative Assessment of Syllabic Timing Deficits in Ataxic Dysarthria.” EMBC, pp. 425–428. IEEE, 2018.

[2]. Rusz Jan, Benova Barbora, Ruzickova Hana, Novotny Michal, Tykalova Tereza, Hlavnicka Jan, Uher Tomas et al. “Characteristics of motor speech phenotypes in multiple sclerosis.” Multiple sclerosis and related disorders 19 (2018): 62–69. [PubMed: 29149697]

[3]. Poellabauer Christian, Yadav Nikhil, Daudet Louis, Schneider Sandra L., Busso Carlos, and Flynn Patrick J.. “Challenges in concussion detection using vocal acoustic biomarkers.” IEEE Access 3 (2015): 1143–1160.

[4]. Godino-Llorente JI, Shattuck-Hufnagel S, Choi JY, Moro-Velzquez L, and Gmez-Garca JA. “Towards the identification of Idiopathic Parkinsons Disease from the speech. New articulatory kinetic biomarkers.” PloS one 12, no. 12 (2017): e0189583.

[5]. Duranovic Mirela, and Sehic Sabina. “The speed of articulatory movements involved in speech production in children with dyslexia.” Jour. of learning disabilities 46, no. 3 (2013): 278–286.

[6]. Son Guiyoung, Kwon Soonil, and Lim Yoonseob. “Speech rate control for improving elderly speech recognition of smart devices.” AECE, no. 2 (2017): 79–85.

[7]. Zhang Hanbin, Wang Aosen, Li Dongmei, and Xu Wenyao. “Deep-Voice: A voiceprint-based mobile health framework for Parkinson’s disease identification” BHI, EMBS, pp. 214–217. IEEE, 2018.

[8]. Tao Fei, Daudet Louis, Poellabauer Christian, Schneider Sandra L., and Busso Carlos. “A Portable Automatic PA-TA-KA Syllable Detection System to Derive Biomarkers for Neurological Disorders.” In INTER-SPEECH, pp. 362–366. 2016.

[9]. Pons Jordi, Gong Rong, and Serra Xavier. “Score-informed syllable segmentation for a cappella singing voice with convolutional neural networks.” arXiv preprint arXiv:1707.03544 (2017).

[10]. LeCun Yann, Bengio Yoshua, and Hinton Geoffrey. “Deep learning.” nature 521, no. 7553 (2015): 436.

[11]. Wang Yu-Tsai, Kent Ray D., Duffy Joseph R., and Thomas Jack E.. “Analysis of diadochokinesis in ataxic dysarthria using the motor speech profile program.” Folia Phoniatrica et Logopaedica 61, no. 1 (2009): 1–11.

[12]. Kumar Anurag, and Raj Bhiksha. “Deep cnn framework for audio event recognition using weakly labeled web data.” arXiv preprint arXiv:1707.02530 (2017).

[13]. BenDavid Boaz M., and Icht Michal. “Oraldiadochokinetic rates for Hebrewspeaking healthy ageing population: nonword versus realword repetition.” Int. jour. of language communication disorders 52, no. 3 (2017): 301–310.

[14]. Rsnen Okko, Doyle Gabriel, and Frank Michael C.. “Pre-linguistic segmentation of speech into syllable-like units.” Cognition 171 (2018): 130–150. [PubMed: 29156241]

[15]. Smekal Zdenek, Mekyska Jiri, Rektorova Irena, and Faundez-Zanuy Marcos. “Analysis of neurological disorders based on digital processing of speech and handwritten text” ISSCS, pp. 1–6. IEEE, 2013.

[16]. Mekyska J, Smekal Z, Kostalova M, Mrackova M, Skutilova S, and Rektorova I. “Motor aspects of speech imparment in Parkinson’s disease and their assessment.” CESKA A SLOVENSKA NEUROLOGIE A NEUROCHIRURGIE 74, no. 6 (2011): 662–668.

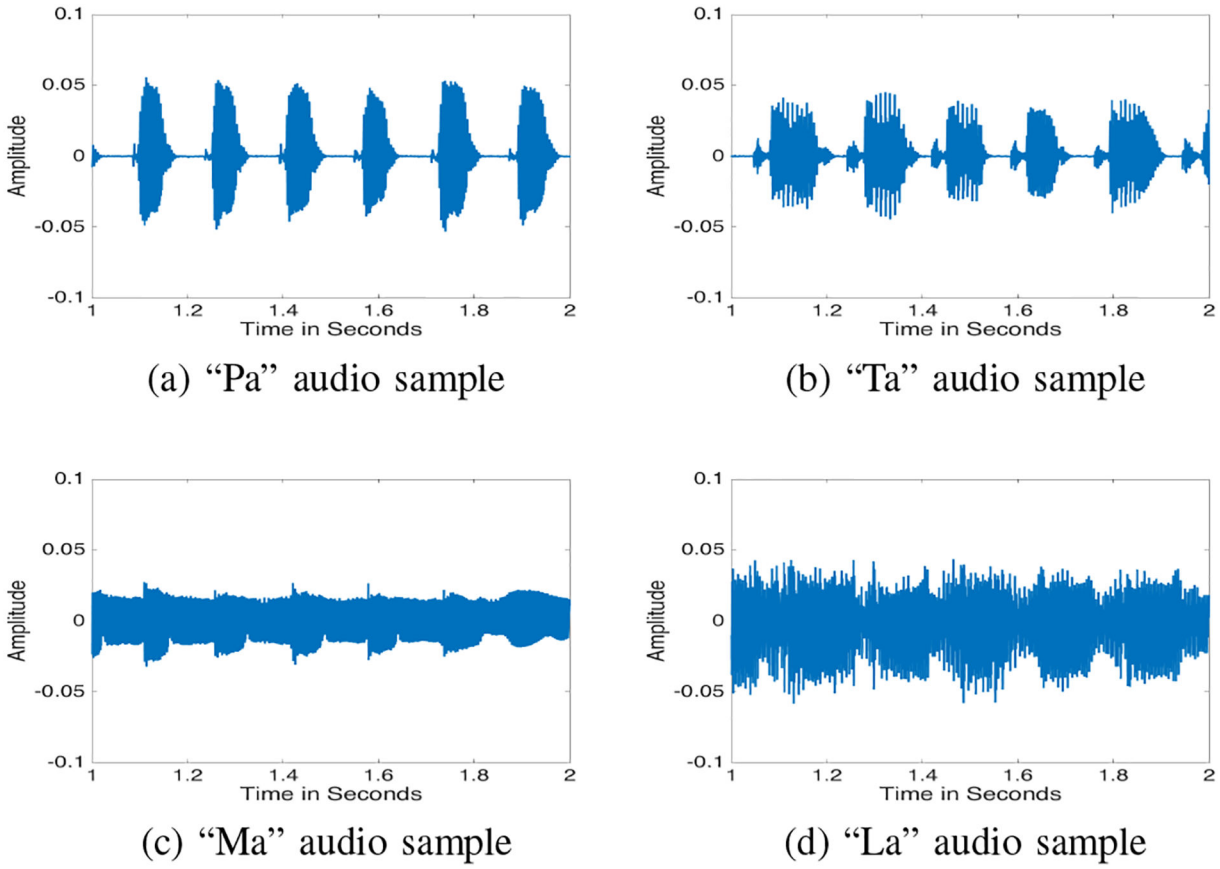


Fig. 1:
Audio waveform samples for different types of oral-DDK tasks.

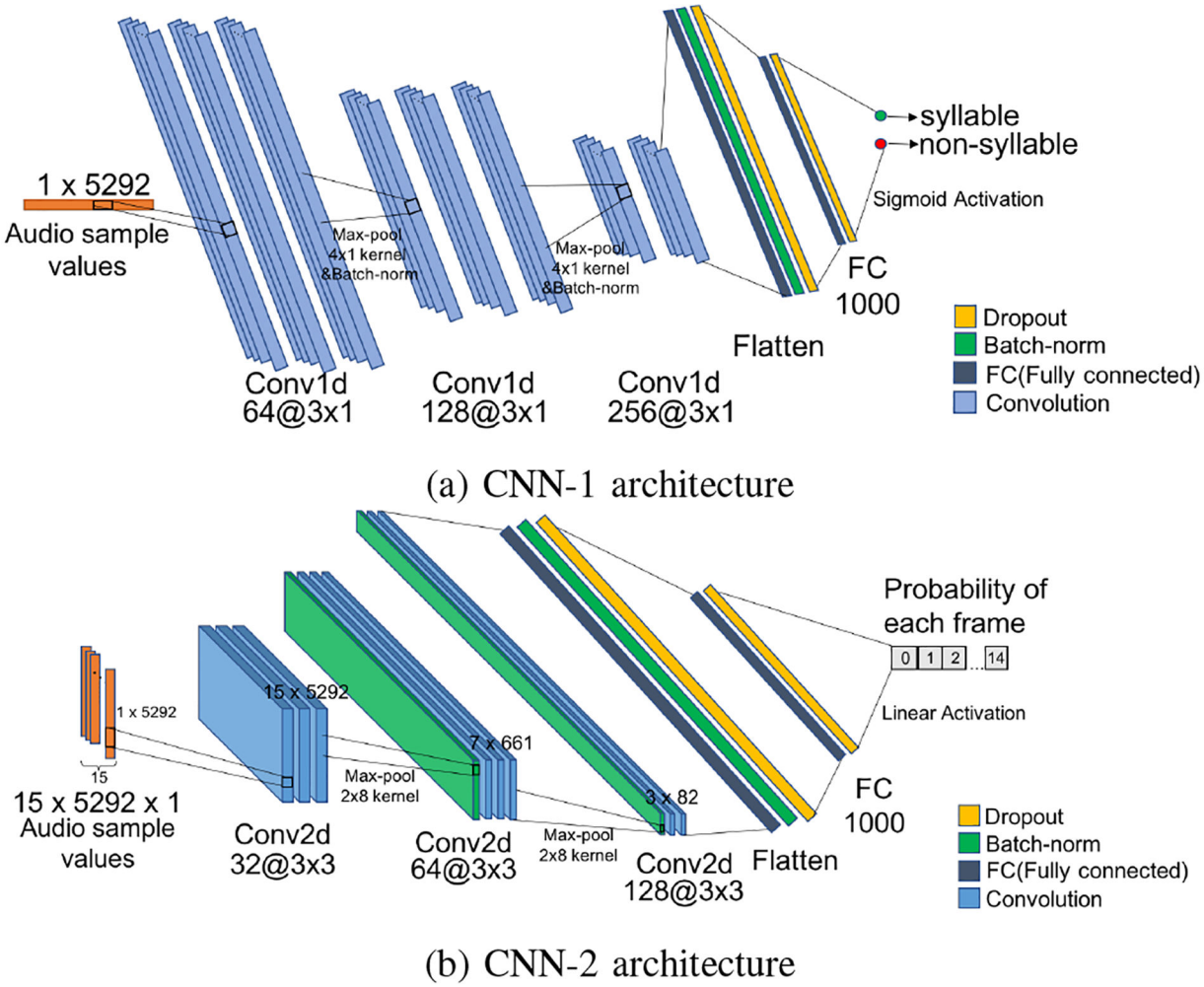
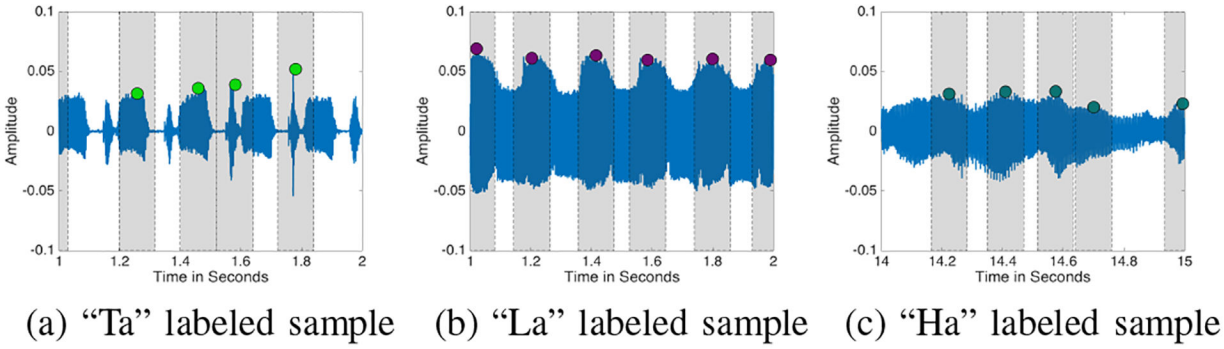


Fig. 2:
CNN-1 and CNN-2 architectures used for DDK syllable detection and localization.



(a) "Ta" labeled sample (b) "La" labeled sample (c) "Ha" labeled sample

Fig. 3:

Sample signals and associated training data. Colored dots mark ground-truth timestamps, shaded regions mark positive training samples for CNN-1.

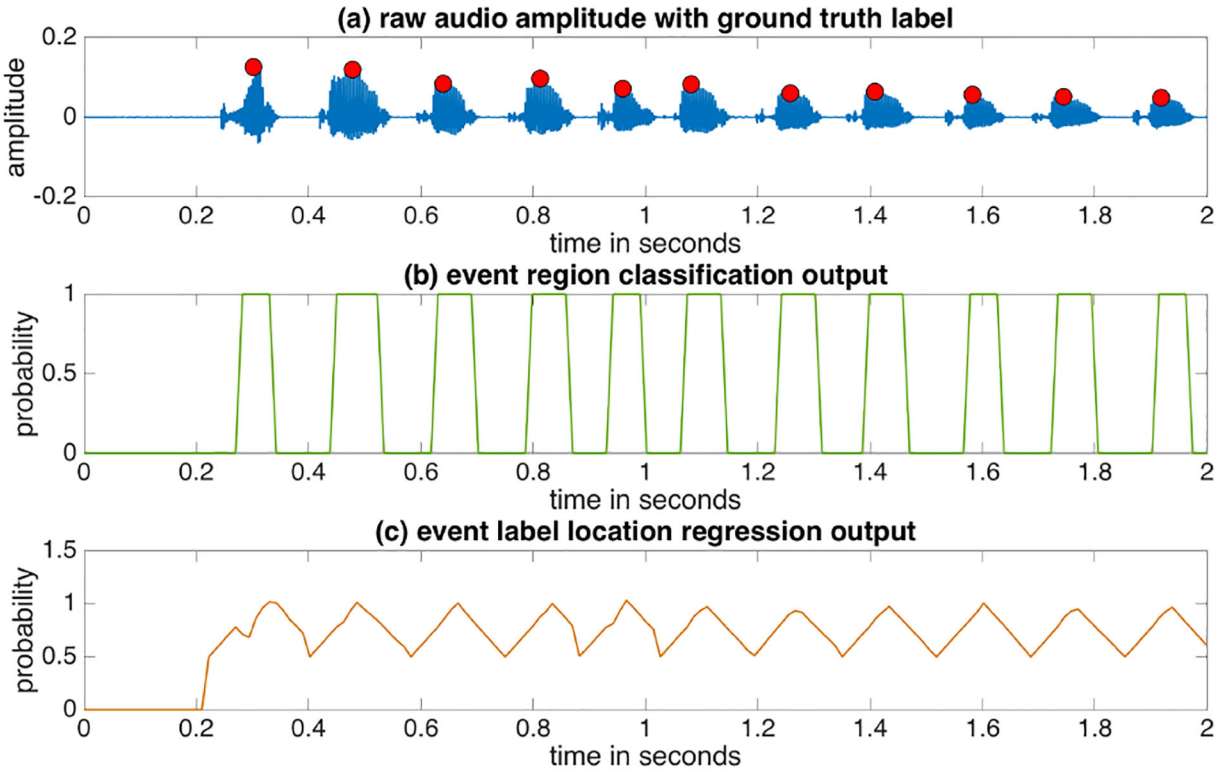


Fig. 4:

Intermediate outputs from the different stages of DeepDDK for a sample “Pa” file. Top panel: original audio signal (blue) with ground-truth timestamps (red). Second panel: output of CNN-1. Third panel: output of CNN-2 where local maxima indicate syllable timestamp.

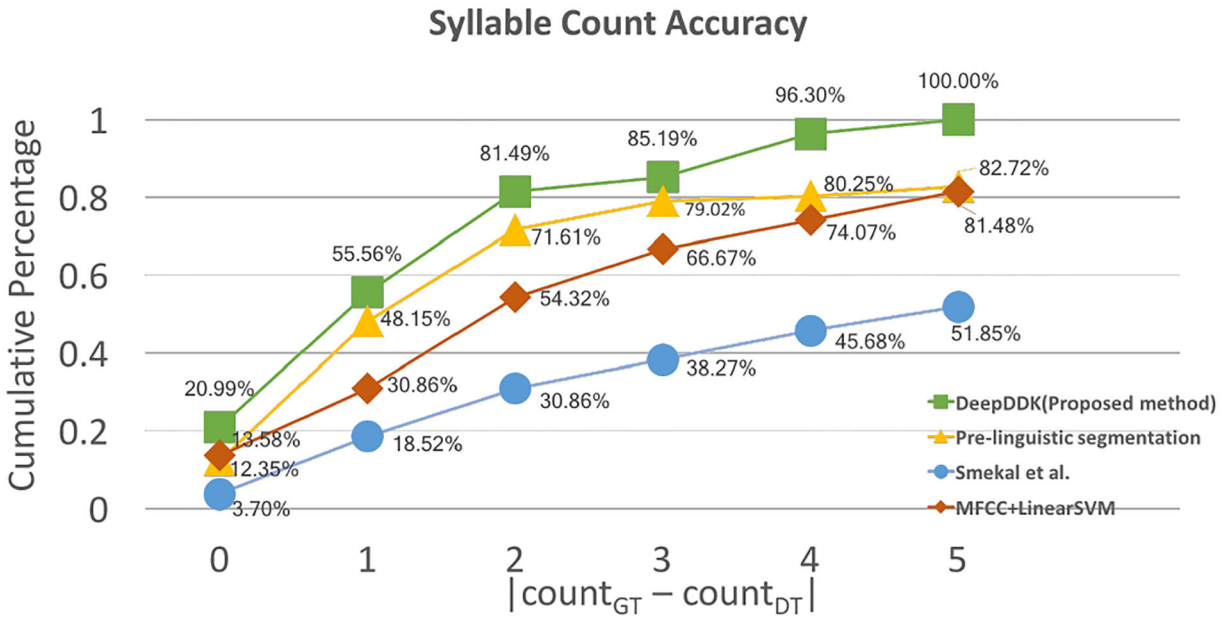


Fig. 5:

Cumulative distribution of event count error for pre-linguistic segmentation [14], Smekal et al. [15] [16], MFCC with Linear SVM and our DeepDDK software. Horizontal axis indicates count error (difference between the number of predicted events vs. ground truth events). Vertical axis shows the ratio of the test files. Absolute event count differences of 1, 2, 3, 4, 5 in the graph correspond to percent count errors of 1.35%, 2.70%, 4.05%, 5.40%, 6.75%, respectively (average number of events per file is 74).

TABLE I:

Layer details for CNN-1 and CNN-2 used for DDK syllable detection and localization.

CNN-1: Layer (type)	Size	CNN-2: Layer (type)	Size
input_l (InputLayer)	(5292, 1)	input_2 (InputLayer)	(15, 5292, 1)
convl (ConvlD)	(5292, 64)	convl (Conv2D)	(15, 5292, 32)
conv2 (ConvlD)	(5292, 64)	conv2 (Conv2D)	(15, 5292, 32)
conv3 (ConvlD)	(5292, 64)	conv3 (Conv2D)	(15, 5292, 32)
max_pooling1d_ 1	(1323, 64)	max_pooling2d_ 1	(7, 661, 32)
batch_norm_l	(1323, 64)	batch_norm_5	(7, 661, 32)
conv4 (ConvlD)	(1323, 128)	conv4 (Conv2D)	(7, 661, 64)
conv5 (ConvlD)	(1323, 128)	conv5 (Conv2D)	(7, 661, 64)
conv6 (ConvlD)	(1323, 128)	conv6 (Conv2D)	(7, 661, 64)
max_poolingld_2	(330, 128)	max_pooling2d_2	(3, 82, 64)
batch_norm_2	(330, 128)	batch_norm_6	(3, 82, 64)
conv7 (ConvlD)	(330, 256)	conv7 (Conv2D)	(3, 82, 128)
conv8 (ConvlD)	(330, 256)	conv8 (Conv2D)	(3, 82, 128)
flatten_l (FC)	(84480)	flatten_2 (FC)	(31488)
batch_norm_3	(84480)	batch_norm_7	(31488)
dropout_l	(84480)	dropout_3	(31488)
dense_l (FC)	(1000)	dense_3 (FC)	(1000)
dropout_2	(1000)	dropout_4	(1000)
dense_2 (FC)	(2)	dense_4 (FC)	(15)
batch_norm_4	(2)	batch_norm_8	(15)
output_classiflcation	(2)	output_regression	(15)

TABLE II:

DeepDDK's location accuracy of different types of syllables.

Type	Event Duration	Recall		Precision	
		30ms	60ms	30ms	60ms
'Pa'	120ms	0.97	0.98	0.97	0.98
'Ta'	170ms	0.81	0.95	0.81	0.95
'Ka'	140ms	0.91	0.97	0.92	0.98
'Ba'	90ms	0.97	0.98	0.97	0.99
'Da'	130ms	0.89	0.97	0.81	0.98
'Ga'	110ms	0.94	0.96	0.95	0.98
'La'	100ms	0.79	0.90	0.79	0.90
'Ma'	90ms	0.88	0.95	0.89	0.97
'Ha'	140ms	0.85	0.93	0.87	0.95
Average	120ms	0.89	0.95	0.90	0.97