LRDWWS CHALLENGE SYSTEM DESCRIPTION - TEAM T019

Yan Xiong[1], *Si-Ioi Ng*[2]

[1]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, USA
[2]College of Health Solutions, Arizona State University, Tempe, USA

ABSTRACT

This technical report presents our proposed solution to developing a wake-up word spotting system for the SLT 2024 Low-Resource Dysarthria Wake-Up Word Spotting Challenge (LRDWWS). Under the low-resource setting, we are focused on 1) the use of pre-trained self-supervised learning (SSL) speech foundation model for front-end acoustic feature extraction, 2) the use contrastive learning to enhance neural network embeddings for wake-word detection, and 3) the use of clinically-inspired voice and prosodic features to improve the model's adaption to dysarthric speakers.

1. FEATURE EXTRACTION WITH WAV2VEC2

In the baseline method, the filterbank (FBank) features are computed from raw speech samples and are used as the input to the backbone feature extractor. In our system, we replace the FBank feature computation with a pre-trained Wav2Vec2 model [1]. The Wav2vec2 model is a self-supervised learning framework for automatic speech recognition that learns speech representations from raw audio data. We use the "wav2vec2-large-xlsr-53" pre-trained model provided on HuggingFace [2], whose training set includes Mandarin samples. The hidden states of the Wav2vec2 model are used as the input to the CNN backbone. Table 1 shows the performance of using Wav2Vec2 on the enrollment dataset. The comparison shows that the model results in lower error rates for all 4 speakers when using Wav2vec2 features.

Table 1. Feature Computation: FBank vs. Wav2Vec2.

Speaker	Intelligibility	FBank Feature	Wav2Vec2
DF0015	68.44	0.233	0.138
DF0016	93.73	0.077	0.046
DM0005	85.78	0.243	0.132
DM0019	47.95	0.391	0.297

Given the control (con.) and uncontrol (uncon.) datasets, the baseline method divides the model pre-training into 2 stages where the first stage trains the model with control samples and the second stage uses the uncontrol samples. In our design, instead of using only uncontrol data, we use both control and uncontrol data so that the system uses control data for regularization. Table 2 shows the performance of using both control and uncontrol data in stage 2 on the enrollment dataset. The results show that using both control and uncontrol data improves the performance on 3 out of the 4 speakers.

Table 2. Stage 2 Data Usage

Speaker	Intelligibility	uncon.	uncon.+con.
DF0015	68.44	0.138	0.102
DF0016	93.73	0.046	0.077
DM0005	85.78	0.132	0.103
DM0019	47.95	0.297	0.258

2. CONTRASTIVE LEARNING TO ENHANCE DYSARTHRIC WAKE WORD SPOTTING

We consider dysarthria as a form of distortion to the speech, and we want to train the feature extractor to "enhance" the dysarthric speech. To achieve this goal, we design a contrastive learning strategy to make the feature extraction backbone network learn from the similarity and divergence between keywords from control and uncontrol speakers. In contrastive learning, each keyword sample is paired with another keyword sample from speakers from different classes. For example, when the sample is from a control speaker, then the other sample comes from an uncontrol speaker. As is shown in figure 1, both samples are processed by Wav2Vec2 model and feature extractor to generate the feature vectors.

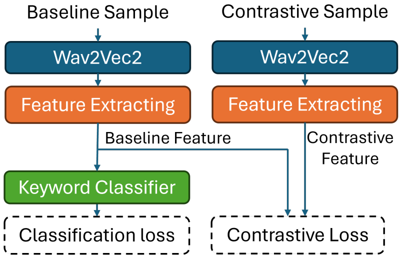

Fig. 1. Data flow of Contrastive Learning

A contrastive loss function is computed using the Euclidean distance between the two feature vectors as in shown in eq 1. L denotes the contrastive label with a value of 1 indicating the samples are from same keyword class and a value of 0 indicating the samples are from different keyword classes. The vectors f_1 and f_2 denote the baseline feature vector and contrastive feature vector, and $D(f_1, f_2)$ denotes the euclidean distance. The value m is the margin in contrastive loss, and we use $m = 1$ during contrastive training.

$$\begin{aligned} L_{contrastive} =& L * D(f_1, f_2)^2 + \\ & (1 - L) * max(m - D(f_1, f_2), 0)^2 \end{aligned} \tag{1}$$

As a result, the contrastive loss drives the divergence between two feature vectors to be small when they are from same keyword, or increase the divergence up to the value of the margin when they are from different keywords. We use the contrastive learning loss as a regularization term and add it to the cross-entropy loss with a fixed weight λ. The overall loss function becomes eq 2. We evaluated four cases where λ is set at 0.1, 0.005, 0.003 and 0.001. The results are shown in table 3. The results show that when $\lambda = 0.005$, all speakers in the enrollment show lower error scores.

$$Loss = L_{cross-entropy} + \lambda L_{contrastive} \tag{2}$$

Table 3. Contrastive Learning Performance

Speaker	**Baseline**	**λ Value**			
		0.01	0.005	0.003	0.001
DF0015	0.102	0.106	0.100	0.088	0.111
DF10016	0.077	0.040	0.034	0.063	0.097
DM0005	0.103	0.133	0.095	0.090	0.108
DM0019	0.258	0.322	0.234	0.275	0.310

3. VOICE AND PROSODIC FEATURES

In addition to utilizing deep learning models to learn segmental dysarthric speech characteristics, we also use voice and prosodic speech features in our proposed keyword-spotting system to capture supra-segmental characteristics. Dysarthric patients are known to experience disruptions in controlling their neuromuscular system. This affects their articulation, resonance, phonation, and prosody [3]. The segmental features capture articulation and resonance features. Extraction of the voice and prosodic features is based on signal processing methods and does not require training, which suits low-resource applications, such as the one in this challenge.

We select several voice and prosodic features known to be impacted by dysarthria. The selected features include: 1) speaking rate, 2) mean and standard deviation of fundamental frequency (F0), 3) harmonic-to-noise ratio (HNR), 4) jitter (local, absolute local, relative average perturbation, five-point period perturbation quotient, and difference of differences of periods), 5) shimmer (local, absolute local, three-point, five-point and eleven-point amplitude perturbation quotient, and difference of differences of amplitudes), 6) cepstral peak prominence (CPP), 7) envelope modulated spectrum (EMS) of the 7 octave bands. Speaking rate can serve as a measure of weakness and/or incoordination of the speech musculature. The statistics of F0 describes its longer-term dynamics within an utterance. For example, someone with monotone speech will have a low standard deviation of F0. Shimmer and jitter features represent short-term variations in the F0 contour and are often used to characterize vocal quality. The CPP measures the overall level of noise in the speech signal. It typically describes breathiness in speech and can indicate voice problems [4]. EMS is a representation of the slow amplitude modulation signal. It describes the distribution of energy in amplitude fluctuations across different designated frequency bands. In [5], EMS was shown to effective in automatically discriminating between dysarthria subtypes.

Extraction of duration, F0, HNR, jitter and shimmer relies on Parselmouth [6]. The code of extraction is adopted from an open-sourced toolkit on GitHub[1]. The implementation of EMS extraction follows the recipe in [5]. The modulation spectra for amplitude envelopes extracted from the full speech signal and the six selected octave bands centered at 125, 250, 500, 1000, 2000 and 4000 Hz. The amplitude envelope of each band is extracted by a 30-Hz low-pass fourth-order Butterworth filter (half-wave rectified), and downsampled to 80 Hz with mean subtraction. The power spectrum of a downsampled envelope is calculated with a 512-point fast Fourier transform. From each of the modulation spectra, the peak frequency, peak amplitude, energy in the region of 3-6 Hz, energy in spectrum from 0-4 Hz (Below4), energy in spectrum from 4-10 Hz (Above4), and the ratio between Below4 and Above4 are computed. The 7 amplitude envelopes deliver a 42-dimensional EMS feature in total. Combining all selected voice and prosodic features with a binary gender label, a 59-dimensional feature is constructed and concatenated with the neural network embedding to perform the keyword classification tasks. Table 4 reports the results of using voice and prosodic features in comparison with system not using them.

Table 4. Model performance with and without using voice and prosodic features.

Speaker	**Intelligibility**	**with features**	**without feature**
DF0015	68.44	**0.056**	0.102
DF0016	93.73	**0.069**	0.077
DM0005	85.78	0.144	**0.103**
DM0019	47.95	**0.183**	0.258

4. TEST-B SCORES

We combine the methodologies described in Section 1-3 and produce the final system for submission. Table 5 shows the scores of our submission to test-B leaderboard. Contrastive learning was not included in submissions 1-3 but was adopted in submissions 4-6. An universal threshold of 0.005 was chosen for the best performing system.

Table 5. Test-B Submission Scores

Submission #	**Score**	**FAR**	**FRR**
1	0.121	0.032	0.089
2	0.184	0.017	0.167
3	0.117	0.022	0.095
4	0.099	0.020	0.079
5	0.104	0.020	0.084
6	0.109	0.016	0.093

[1] https://github.com/drfeinberg/PraatScripts?tab=readme-ov-file

5. REFERENCES

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[2] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., “Huggingface’s transformers: State-of-the-art natural language processing,” *arXiv preprint arXiv:1910.03771*, 2019.

[3] Pam Enderby, “Disorders of communication: dysarthria,” *Handbook of clinical neurology*, vol. 110, pp. 273–281, 2013.

[4] Olivia Murton, Robert Hillman, and Daryush Mehta, “Cepstral peak prominence values for clinical voice evaluation,” *American Journal of Speech-Language Pathology*, vol. 29, no. 3, pp. 1596–1607, 2020.

[5] Julie M. Liss, Sue LeGendre, and Andrew J. Lotto, “Discriminating dysarthria type from envelope modulation spectra,” *Journal of Speech, Language, and Hearing Research*, vol. 53, no. 5, pp. 1246–1255, 2010.

[6] Yannick Jadoul, Bill Thompson, and Bart De Boer, “Introducing parselmouth: A python interface to praat,” *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.