LEARNING TO DETECT DYSARTHRIA FROM RAW SPEECH

Juliette Millet[1,2], Neil Zeghidour[1,3]

[1] CoML, ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France
[2] Laboratoire de Linguistique Formelle, CNRS/Paris Diderot University/Sorbonne Paris Cité, Paris, France
[3] Facebook A.I. Research, Paris, France

ABSTRACT

Speech classifiers of paralinguistic traits traditionally learn from diverse hand-crafted low-level features, by selecting the relevant information for the task at hand. We explore an alternative to this selection, by learning jointly the classifier, and the feature extraction. Recent work on speech recognition has shown improved performance over speech features by learning from the waveform. We extend this approach to paralinguistic classification and propose a neural network that can learn a filterbank, a normalization factor and a compression power from the raw speech, jointly with the rest of the architecture. We apply this model to dysarthria detection from sentence-level audio recordings. Starting from a strong attention-based baseline on which mel-filterbanks outperform standard low-level descriptors, we show that learning the filters or the normalization and compression improves over fixed features by 10% absolute accuracy. We also observe a gain over OpenSmile features by learning jointly the feature extraction, the normalization, and the compression factor with the architecture. This constitutes a first attempt at learning jointly all these operations from raw audio for a speech classification task.

Index Terms— dysarthria, paralinguistic, classification, waveform, lstm

1. INTRODUCTION

Learning from speech still relies on handcrafted, fixed features on which a classifier can be trained. This differs from a field like computer vision which now widely uses end-to-end models trained on raw pixels, that are typically processed by learnable convolutional operations [1, 2, 3]. Speech features typically contain spectral representations, such as mel-filterbanks or MFCCs, and/or low-level informations [4], such as zero-crossing rate or harmonics-to-noise ratio. They are chosen to model a broad range of linguistic and paralinguistic information. Training a classifier from these fixed coefficients requires performing a feature selection step, which has the limitation that it cannot retrieve useful information that would have been lost in the feature computation. Recent research has shown improvement when replacing fixed speech features by
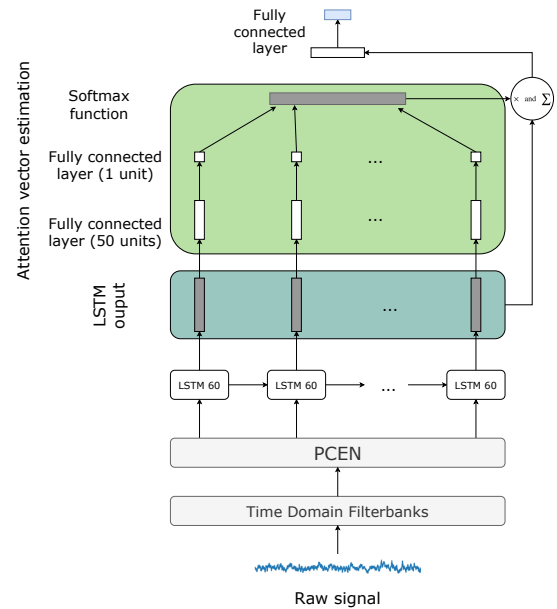

Fig. 1. Proposed pipeline that learns jointly the feature extraction, the compression, the normalization and the classifier.

a learnable frontend, for tasks such as speech recognition [5], speaker identification [6] or emotion recognition [7]. In this work, we propose to apply such end-to-end systems to another paralinguistic task: the detection of dysarthria from speech recordings. There is a growing interest in automatically extracting information from speech for health care [8, 9, 10], and unlike a feature-driven approach that would require testing various combinations of fixed features, we implement a system that can directly process raw speech and learn relevant features jointly with the dysarthria classifier, such that they will be optimal for the task.

The TORGO database [11] is a collection of annotated speech recordings and articulatory measurements from speakers with cerebral palsy (CP) or amyotrophic lateral sclerosis (ALS), as well as control patients. [12, 13, 14] have used this database to provide speech recognition systems with robustness to dysarthria. [15] trains various linear classifiers on

TORGO and the NKI CCRT corpus [16] to detect dysarthria. More recently, [17] has trained fully connected neural networks to classify the severity of the disease, using TORGO and the UASPEECH [18] database. All these models are trained on standard low-level features. In this work we show that dysarthria detection benefits significantly from learning directly from the raw waveform.

Previous work has explored learnable alternatives to speech features that rely on a similar computation to spectral representations [19, 20, 21, 22, 5]. These approaches learn convolutions that are then passed through a non-linearity, eventually a pooling operator and then a log compression to replicate the dynamic range compression typically performed on spectrograms or mel-filterbanks. This compression function remains fixed and is chosen beforehand, which could impact the final performance, as various compression functions including logarithm, cubic root, or $10th$ root have been previously showed to perform better depending on the task (see Table 2 of [23]). A second fixed component is the mean-variance normalization of speech features. [5] integrates this normalization into the neural architecture, but keeps it fixed during training. [24] introduces a computational block, the Per Channel Energy Normalization (PCEN) that can learn a compression and a normalization factor per channel, and can be integrated into a neural network on top of speech features. It has since then been used in production speech recognition systems [25].

In this work, we start from an attention-based model on mel-filterbanks, which already outperforms an equivalent model trained on low-level descriptors (LLDs). Our experiments show that by training a PCEN block on top of mel-filterbanks or replacing them by learnable time-domain filterbanks from [22], we get a gain in accuracy around 10% in absolute when training an identical neural network for dysarthria detection. Finally, by combining time-domain filterbanks and PCEN we propose the first audio frontend that can learn features, compression and normalization jointly with a neural network using backpropagation.

2. MODEL

2.1. Time-Domain filterbanks

As the first step of our computational pipeline, we use Time-Domain filterbanks from [22]. Time-Domain filterbanks are neural network layers that take the raw waveform as input. They can be initialized to replicate mel-filterbanks, and then learnt for the task at hand. The standard computation of mel-filterbanks relies on passing a spectrogram through a bank of frequency domain filters. More formally, the nth mel-filterbank of a signal in t is:

$$M^n x(t) = \frac{1}{2\pi} \int |\hat{x}_t(w)|^2 |\hat{\psi}_n(w)|^2 \, \mathrm{d}w.$$

Layer type	Input	Output	Width	Stride
Conv. 1D	1	128	400	1
Modulus	128	64	-	-
Square	-	-	-	-
Conv. 1D	64	64	400	160
Absolute value	-	-	-	-
Add 1, log	-	-	-	-

Table 1. Description of the neural network layers used to compute time-domain filterbanks. The parameters are chosen to replicate 64 mel-filterbanks of window size 25ms and stride 10ms at 16kHz.

Where $x_t(u) = \phi(t - u)x(u)$ is the waveform windowed with an Hanning function ϕ centered in t, $(\psi_n)_{n=1...N}$ the N melfilters and $\hat{f}$ denotes the Fourier transform of f.

[26] shows that these coefficient can be approximated in the time domain by the following computation, referred as the first order scattering transform:

$$M^n x(t) \approx |x * \varphi_n|^2 * |\phi|^2(t).$$

Where $(\varphi_n)_{n=1...N}$ are Gabor wavelets defined in [22] such that $|\hat{\varphi}_n|^2 \approx |\hat{\psi}_n|^2$. [22] shows that this computation can be implemented as neural network layers, referred as Time-Domain filterbanks (TD-filterbanks). The waveform goes through a complex-valued convolution, a modulus operator and the a convolution with a lowpass-filter (the squared hanning window) that performs the decimation. When not combined with PCEN, a log-compression is added on top of TD-filterbanks after adding 1 to their absolute value to avoid numerical issues. Table 1 shows the detailed layers.

Following [22], the first 1D convolution filters are initialized with Gabor wavelets, to replicate mel-filterbanks, and are then learnt at the same time as the rest of the model. The second convolution layer is kept fixed as a squared hanning window to perform lowpass filtering.

2.2. Per Channel Energy Normalization

Per Channel Energy Normalization (PCEN) is a learnable component introduced in [24] which computes parametrized normalization and compression. It replaces the log-compression and the mean-variance normalization. With $E(t, f)$ the value of the feature f at time t, the computation of PCEN is:

$$PCEN(t, f) = \left(\frac{E(t, f)}{(\epsilon + M(t, f))^\alpha} + \delta\right)^r - \delta^r.$$

$M(t, f)$ is a moving average of the feature f along the time axis, defined as:

$$M(t, f) = (1 - s)M(t - 1, f) + sE(t, f).$$

α controls the strength of the normalization, the exponent r (typically in $[0, 1]$) defines the slope of the compression curve, s sets the spread of the moving average, and ϵ is a small scalar used to avoid division by zero. By backpropagation, we learn α, r, and δ with the rest of the model to obtain a compression and a normalization that fit the task at hand.

2.3. LSTM and Attention model

The output of the learnable frontend is fed to an attention-based model [27], that contains one LSTM layer of hidden size 60 followed by an attention mechanism, inspired by [28]. The attention mechanism consists of two fully connected layers, of 50 and 1 unit respectively, and a softmax layer, that are applied to each output of the LSTM. The vector obtained is used to weight a linear combination of the LSTM outputs, that goes throught another fully connected layer of size the number of labels considered. The detailed architecture is shown in Figure 1. In [28], this model reaches state-of-the-art performance when trained for emotion recognition on mel-filterbanks, which motivated using it for the paralinguistic task of dysarthria detection.

3. EXPERIMENTAL SETUP

We carry experiments on the TORGO database [29]. It consists of sound recordings, sampled at 16kHz, from speakers with either cerebral palsy or amyotrophic lateral sclerosis, which are two of the prevalent causes of speech disability or dysarthria. Similar data for a control set of subjects is also available. Along with sound recordings, TORGO contains 3D articulatory features that we did not use.

There are five groups of people: the control group not affected by the disease, and 4 other groups of affected people, classified by the severity of the disease. Each person recorded has a code name, **F** is for female, **M** is for male, while **C** is for control, followed by an identification number. A random split of the database would result in similar speakers in training, validation, and test sets, that could reduce the task to a speaker identification task. To avoid this confounding factor, we split the database to have a good repartition of the different severities among the training, validation and test set, while having no common speakers between the different sets (see Table 2 for the detailed split).

After studying the database we decided to pad the recordings so they all last 2.5s. We extracted some typical low level descriptors (LLDs) from it to have a first baseline. We use the OpenSmile toolkit [4], with the configuration of the Interspeech 2009 Emotion Challenge [30]. For each 25ms window of the recordings (strided by 10ms), 32 features are extracted (12 MFCCs, root mean square energy, zero-crossing rate, harmonics-to-noise ratio, F_0 and their Δ).

Our second baseline takes as input mel-filterbanks. We pre-emphase the sound signals with a factor of 0.97. 64 mel-

Training	Validation	Test
FC02, F03 (VL), F01 (L), MC04, MC03, M02 (M)	MC02, FC01, M03 (VL), M01 (M)	FC03, F04 (VL), MC01, M05 (L), M04 (M)
3182 C, 1382 D	950 C, 802 D	2103 C, 997 D

Table 2. Speakers and number of recordings per set: C is control and D is dysarthric, the severity of each person is indicated after their ID: VL is Very Low, L is Low, and M is Medium.

Input data	UAR % val.	UAR % test
LLDs	64.8 ± 1.2	65.5 ± 3.6
mel-filterbanks	79.9 ± 6.3	72.4 ± 3.0
mel-filterbanks + mvn	63.5 ± 1.7	70.3 ± 2.9
mel-filterbanks + PCEN	76.0 ± 6.1	79.7 ± 3.8
Time-Domain filterbanks	93.7 ± 1.2	82.4 ± 0.4

Table 3. UAR (%) of the attention-based model trained over different features or learnable frontends. The UAR is averaged over 3 runs and standard deviations are reported.

filterbanks are computed every 25ms with a stride of 10ms and passed through a log-compression. To evaluate our learnable frontend in a comparable setting, we design them with the same number of filters, window size and stride (see Table 1).

For the PCEN layer, we take $\epsilon = 10^{-6}$ and $s = 0.5$, both fixed, and we only consider the absolute value of r. We initialize r, α and δ at 0.5, 0.98 and 2.0 respectively. All models are trained with a stochastic gradient descent with momentum (0.98) and batch size 1, with a learning rate of 0.001.

We use the Unweighted Average Recall (UAR) to evaluate our results. The UAR of a model is the mean of its accuracy for each label. It is a better metric, when dealing with unbalanced datasets, than the accuracy, since it is reweighting the results depending on the size of each class. It has been widely used in unbalanced settings such as the Emotion Recognition challenge [30]. We use the validation set for hyperparameter selection and early stopping.

4. RESULTS

Table 3 shows the UAR on the validation and test sets. All the results are the mean UAR obtained over three runs with different random initialization. We do not compare them to previously published results [15, 17] as they use additional data and/or perform a different task. The attention based-model trained on LLDs features reaches an accuracy of 66% and is our baseline system. Replacing LLDs by mel-filterbanks improves the performance by 6% in absolute. Adding a fixed mean-variance normalization step (mvn) brings the models to

Input data	UAR % val.	UAR % test
TD-f + PCEN	72.3 ±1.5	74.8 ±1.1
TD-f + PCEN only r	74.6 ±2.9	76.4 ±1.8
TD-f + PCEN only α	66.6 ±1.4	63.3 ±8.2

Table 4. UAR (%) of the attention-based model trained over different fully learnable frontends. The UAR is averaged over 3 runs and standard deviations are reported.

over-fitting, and thus the UAR decreases of 2%. However, we observe that replacing the fixed log-compression and mean-variance normalization step by a learnable PCEN layer improves the UAR of the models of 7% compared to the unnormalized mel-filterbanks. Moreover, an even bigger increase is noticed when replacing mel-filterbanks by equivalent TD-filterbanks (10% in absolute). We can emphasize the fact that using the TD-filterbanks also leads to a more stable learning process, as the standard deviation along different runs is considerably lower.

When studying the new scale learned by the TD-filterbanks (see Figure 2) we notice that the filters tend to focus around $2000Hz$ and $6500Hz$, which suggests that either those frequencies are crucial to identify dysarthria, or the model might exploit a bias in the dataset. In Figure 3 we observe that the parameters learned by the PCEN layer reproduce similar schemes from one model to another, and that the learnt compression varies between filters, unlike a log-compression which is applied equivalently to all channels.

4.1. Fully learnable frontend

As we observe independent gains from either learning the features or learning the compression-normalization, we explore in our final experiments learning jointly all these operations. We remove the log-compression step of Time-Domain filterbanks and replace it by a PCEN layer. We use three settings: one for which r, α and δ are learned, the second one with only r learned, and finally the last one for which only α is learned. If a parameter is not learned, it is fixed to its initial value (specified in Section 3). Table 4 shows that learning only the normalization exponent gives worse results than the models trained on LLDs. However, we notice that the model learning r, α and δ, and the one only learning r match the models using mel-filterbanks.

5. CONCLUSION

This paper presents a fully learnable audio frontend, combining Time-Domain filterbanks and Per Channel Energy Normalization. It is the first time that a model is developed with the ability to learn the extraction, compression and normalization of the features from the raw waveform, jointly with a classifier. We apply it to dysarthria detection, and show
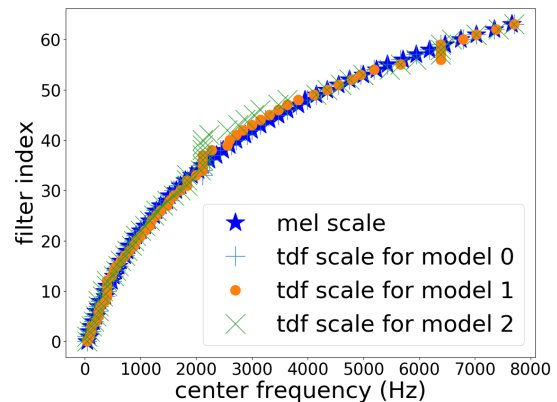

Fig. 2. New scales obtained by three independent models using TD-filterbanks, compared to mel scale. The center frequency is the frequency for which a filter is maximum.
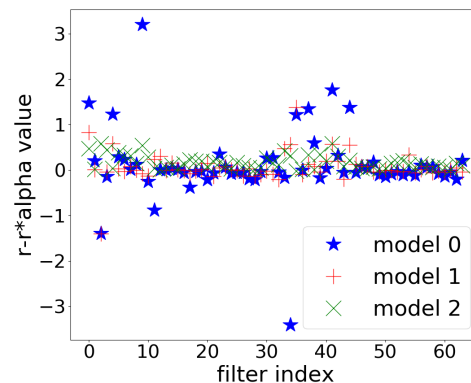

Fig. 3. Approximation of the compression exponent obtained for the PCEN layer learned on mel-filterbanks.

that replacing fixed features by learnable frontends leads to an increase in performance of the models for this task, consistently with previous results on other linguistic and paralinguistic tasks. Learning only the Time-Domain filterbanks or the PCEN parameters gives better results than learning them jointly, but learning both still gives similar to better performance than using fixed features, which constitutes a proof of concept for fully learnable audio frontends.

6. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection

and semantic segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

[3] Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 1915–1929, 2013.

[4] Florian Eyben, Martin Wöllmer, and Björn W. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, 2010.

[5] Neil Zeghidour, Nicolas Usunier, Gabriel Synnaeve, Ronan Collobert, and Emmanuel Dupoux, "End-to-end speech recognition from the raw waveform," in *Interspeech*, 2018.

[6] Hannah Muckenhirn, Mathew Magimai-Doss, and Sébastien Marcel, "Towards directly modeling raw speech signal for speaker verification using cnns," *ICASSP*, pp. 4884–4888, 2018.

[7] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A. Nicolaou, Björn W. Schuller, and Stefanos Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," *ICASSP*, pp. 5200–5204, 2016.

[8] Maree Johnson, Samuel Lapkin, Vanessa Long, Paula Sanchez, Hanna Suominen, Jim Basilakis, and Linda Dawson, "A systematic review of speech recognition technology in health care," in *BMC Med. Inf. & Decision Making*, 2014.

[9] Max A. Little, Patrick E. McSharry, Eric J. Hunter, Jennifer L. Spielman, and Lorraine O. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, pp. 1015–1022, 2009.

[10] Björn W. Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus R. Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, Marcello Mortillaro, Hugues Salamin, Anna Polychroniou, Fabio Valente, and Samuel Kim, "The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism," in *INTERSPEECH*, 2013.

[11] Frank Rudzicz, Aravind Kumar Namasivayam, and Talya Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, Dec 2012.

[12] Kinfe Mengistu and Frank Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *ICASSP*, 05 2011, pp. 4924–4927.

[13] Kinfe Mengistu and Frank Rudzicz, "Comparing humans and automatic speech recognition systems in recognizing dysarthric speech," 05 2011, vol. 6657, pp. 291–300.

[14] Myungjong Kim, J Yoo, and H Kim, "Dysarthric speech recognition using dysarthria-severity-dependent and speaker-adaptive models," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTER-SPEECH*, 08 2013, pp. 3622–3626.

[15] Jangwon Kim, Naveen Kumar, Andreas Tsiartas, Ming Li, and Shrikanth S. Narayanan, "Automatic intelligibility classification of sentence-level pathological speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 132 – 144, 2015.

[16] Renee Peje Clapham, Lisette van der Molen, R. J. J. H. van Son, Michiel W. M. van den Brekel, and Frans J. M. Hilgers, "Nki-ccrt corpus - speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy," in *LREC*, 2012.

[17] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *ICASSP*, March 2017, pp. 5070–5074.

[18] Heejin Kim, Mark Hasegawa-Johnson, Adrienne Perlman, Jon Gunderson, Thomas S. Huang, Kenneth Watkin, and Simone Frame, "Dysarthric speech database for universal access research," in *INTERSPEECH*, 2008.

[19] Yedid Hoshen, Ron J Weiss, and Kevin W Wilson, "Speech acoustic modeling from raw multichannel waveforms," in *Proceedings of ICASSP*. IEEE, 2015.

[20] Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Interspeech*, 2015.

[21] Pegah Ghahremani, Vimal Manohar, Daniel Povey, and Sanjeev Khudanpur, "Acoustic modelling from the signal domain using cnns," in *INTERSPEECH*, 2016.

[22] Neil Zeghidour, Nicolas Usunier, Iasonas Kokkinos, Thomas Schatz, Gabriel Synnaeve, and Emmanuel Dupoux, "Learning filterbanks from raw speech for phone recognition," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5509–5513, 2018.

[23] Ralf Schlüter, Ilja Bezrukov, Hermann Wagner, and Hermann Ney, "Gammatone features and feature combination for large vocabulary speech recognition," *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, vol. 4, pp. IV–649–IV–652, 2007.

[24] Yuxuan Wang, Pascal Getreuer, Thad Hughes, Richard F Lyon, and Rif A Saurous, "Trainable frontend for robust and far-field keyword spotting," in *ICASSP*. IEEE, 2017, pp. 5670–5674.

[25] Eric Battenberg, Rewon Child, Adam Coates, Christopher Fougner, Yashesh Gaur, Jiaji Huang, Heewoo Jun, Ajay Kannan, Markus Kliegl, Atul Kumar, et al., "Reducing bias in production speech models," *arXiv preprint arXiv:1705.04400*, 2017.

[26] Joakim Andén and Stéphane Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, vol. 62, pp. 4114–4128, 2014.

[27] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.

[28] P. Hsiao and C. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *ICASSP*, April 2018, pp. 2526–2530.

[29] Frank Rudzicz, Pascal Van Lieshout, Graeme Hirst, Gerald Penn, Fraser Shein, and Talya Wolff, "Towards a comparative database of dysarthric articulation," in *Proceedings of ISSP*, 01 2008.

[30] Björn Schuller, Stefan Steidl, and Anton Batliner, "The interspeech 2009 emotion challenge," in *Proc. Interspeech*, 01 2009, pp. 312–315.