SPEECH RECOGNITION SYSTEM: EMPLOYABILITY OF CNN IN MITIGATING OVERLAPPING SPEECH/NOISE RESOLUTIONS

Karan Sablok

Student, Delhi Public School

ABSTRACT

Voice detection systems typically models the relationship between the audio voice sign and the phones in two distinct phase: feature mining and classifier learning. In our latest research, we have displayed that, in the framework of CNN, the correlation among the raw voice sign and the phones can be directly demonstrated and ASR systems inexpensive to normal method can be built. In this paper, we first examine and show that, among the first two convolutional layers, the CNN learns (in parts) and models the phone-specific spectral cover info of 2-4 ms voice. Given that we show that the CNN-based method produces ASR styles like to normal temporarily spectral based ASR system under mismatched (noisy) situations, with the CNN-based method being more robust.

Index Terms*: automatic speech recognition, convolutional neural networks, raw signal, robust speech recognition.*

1. INTRODUCTION

Best in class programmed discourse acknowledgment (ASR) frameworks regularly show the connection between the acoustic discourse flag and the telephones in two separate advances, which are streamlined in a free way [1]. In an initial step, the discourse flag is changed into highlights, typically made out of a dimensionality decrease stage and a data determination stage, in view of the undertaking particular learning of the marvels. These two stages have been painstakingly hand-made, prompting best in class highlights, for example, Mel recurrence cepstral coefficients (MFCCs) or perceptual direct forecast cepstral highlights (PLPs). In a second step, the probability of sub word units, for example, phonemes is evaluated utilizing generative models or discriminative models.

As of late, in the half and half HMM/ANN structure [1], there has been developing interests in utilizing "middle of the road" portrayals rather than ordinary highlights, for example, cepstral based highlights, as contribution for neural systems based frameworks. ANNs with profound learning structures, all the more exactly, profound neural systems (DNNs) [2, 3], which can yield preferable framework over a solitary shrouded layer MLP have been proposed to address

different parts of acoustic displaying. All the more particularly, utilization of setting subordinate phonemes [4, 5]; utilization of otherworldly highlights rather than cepstral highlights [6, 7]; CNN-based framework with Mel channel bank energies as info [8, 9, 10]; blend of various highlights [11], to give some examples. Highlights gaining from the crude discourse flag utilizing neural systems based frameworks has likewise been researched in [12]. In every one of these methodologies, the highlights extraction step and the acoustic demonstrating step are prepared freely. All the more as of late, neural system based frameworks where the highlights and the model are prepared together have been proposed. CNN-based framework taking force range as information has been proposed in [13]. Utilizing transient crude discourse straightforwardly as information has been proposed with regards to DNNs [14] and with regards to end-to-end arrangement discriminative preparing of CNNs [15]. In our ongoing investigations [16, 17], it was demonstrated that it is conceivable to evaluate phoneme class restrictive probabilities by utilizing crude discourse motion as contribution to convolutional neural systems [18] (CNNs). On phoneme acknowledgment errand and on ceaseless discourse acknowledgment assignment, we demonstrated that the framework can take in highlights from the crude discourse flag, and yields execution comparative or superior to regular ANN-based framework that takes cepstral includes as information. We likewise demonstrated that the principal convolutional layer of the system can be viewed as an arrangement of coordinating channels, handling the discourse motion at a sub segmental level, 2-4 ms discourse. We demonstrated that these channels react to various recurrence transmission capacities [16], and that they demonstrate some level of invariance crosswise over databases [17]. In this paper, we initially break down the CNN to comprehend the discourse data that is demonstrated between the initial two convolution layers. With that in mind, we present a strategy to figure the mean recurrence reactions of the channels in the primary convolution layer that match to the particular sources of info speaking to vowels. Our investigations on TIMIT assignment demonstrate that the mean recurrence reaction tends to display the envelope of the sub-segmental (2-4 ms) discourse flag. We at that point present an investigation to assess the helplessness of the CNN-based framework to bungled conditions. This is an open issue in frameworks prepared in information driven way. We explore this viewpoint on two errands, to be specific, TIMIT phoneme acknowledgment assignment and Aurora2 associated word acknowledgment undertaking. Our examinations demonstrate that the execution of the CNN based framework corrupts with the reduction in flag to-clamor proportion (SNR) like in a standard unearthly component based framework. Be that as it may, when contrasted with the otherworldly component based framework, the CNN-based framework utilizing crude discourse motion as info yields better execution. The rest of the paper is sorted out as pursues. Area 2 shows the design of the system. Segment 3 introduces the trial setup. Area 4 shows the system investigation and Section 5 exhibits the clamor consider. At long last, Section 6 condenses and finishes up the paper.

2. CONVOLUTIONAL NEURAL NETWORKS

We present quickly the design of the CNN-based framework. More points of interest can be found in [17].

2.1. Design

The convolutional neural system is given a succession of crude info flag, split into casings, and yields a score for every class, for each casing. The system engineering is made out of a few channel stages, trailed by a grouping stage. A channel organize includes a convolutional layer, trailed by a worldly max-pooling layer and a non-linearity (tanh()). Our ideal engineering included three channel stages. Handled signs leaving these stages are encouraged to an arrangement organize, which for our situation is a multi-layer perceptron, with one shrouded layer. It yields the restrictive probabilities p(ijx) for each class I, for each casing x utilizing a SoftMax layer [19]. The system is prepared under the cross-entropy measure, augmented utilizing the stochastic angle rising calculation [20].

2.2. Convolutional layer

While "established" straight layers in standard MLPs acknowledge a fixed size input vector, a convolution layer is thought to be encouraged with a succession of T vectors/outlines: X = fx1 x2 : xT g. A convolutional layer applies the equivalent direct change over each progressive (or interspaced by dW outlines) windows of kW outlines. For instance, the change at edge t is formally composed as:

$$M \begin{pmatrix} x^{t-(kW-1)/2} \\ \vdots \\ x^{t+(kW-1)/2} \end{pmatrix}$$

Where M is a dout x clamor framework of parameters. As it were, doubt channels (lines of the lattice M) are connected to the info arrangement.

2.3. Max-pooling layer

These kind of layers perform local temporal max operations over an input sequence. More formally, the transformation at frame t is written as:

$$\max_{t-(kW-1)/2 \leq s \leq t+(kW-1)/2} x_s^d \qquad \forall d$$

with x being the input, kW the kernel width and d the dimension.

3. EXPERIMENTAL SETUP

3.1. Databases

The TIMIT acoustic-phonetic corpus comprises of 3,696 preparing expressions (examined at 16kHz) from 462 speakers, barring the SA sentences. The cross-approval set comprises of 400 articulations from 50 speakers. The center test set is utilized to report the outcomes. It contains 192 articulations from 24 speakers, barring the approval set. The 61 hand named phonetic images are mapped to 39 phonemes with an extra waste class, as introduced in [21]. For the commotion considers, the articulations from the TIMIT corpus are adulterated by clamors from the NoiseX-92 corpus [22]. The center test set is tainted with the discourse, F-16 and production line clamors, at SNR level somewhere in the range of 0dB and 30 dB. We likewise present a multi-contingent preparing study, where the train set is arbitrarily part in 20 subsets, every one containing 184 expressions. The 20 subsets speak to 4 clamor composes (auto, task, lynx, minigun) not quite the same as the test set, at 5 distinctive SNRs (20dB, 15dB, 10dB, 5dB and clean). The debased expressions are acquired utilizing the FaNT device [23]. The Aurora2 corpus [24] is an associated digit corpus which contains 8,440 sentences of clean and multi-condition preparing information and 70,070 sentences of spotless and uproarious test information. We report the outcomes on test An and test B, made out of 10 unique clamors at 7 diverse commotion levels (perfect, 20dB, 15dB, 10dB, 5dB, 0dB, - 5dB), totaling 70 distinctive test situations, each containing 1,001 sentences. The arrangement is gotten utilizing. the HTK-based HMM/GMM framework furnished alongside the database. It comprises of entire word HMM models with 16 states for every word to show the digits. The states are associated in a basic left-to-right design. The quantity of state is 179. The dialect demonstrates given by the corpus is utilized.

3.2. Tasks

For the associated word acknowledgment undertaking on the Aurora2 corpus, the CNN-based framework is utilized to process the back probabilities of word states. The decoder is a HMM, displaying words. The scaled probabilities are assessed by isolating the back likelihood by the earlier likelihood of each class, evaluated by depending on the preparation set. The hyper parameters, for example, dialect scaling factor and the word inclusion punishment are resolved on the approval set. For the phoneme acknowledgment errand on the TIMIT corpus, the CNN-based framework is utilized to assess phoneme class contingent probabilities. The decoder is a standard HMM decoder, with obliged term of 3 states, and considering all phoneme similarly plausible. We don't utilize a phonetic dialect show. 39 classes are utilized.

3.3. Features Input

Raw features are basically made out of a window of the worldly discourse flag (subsequently, racket = 1 for the principal convolutional layer). The window is standardized with the end goal that it has zero mean and unit difference. We additionally performed pattern tries different things with MFCC as information highlights. They are processed (with HTK [25]) utilizing a 25 ms Hamming window on the discourse motion, with a move of 10 ms. The flag is spoken to utilizing twelfth request coefficients (without the zeroth coefficient) and the logarithmic edge vitality, alongside their first and second subordinates, figured on a 9 outlines setting.

3.4. Baseline systems

We match our methodology and the standard HMM/ANN framework utilizing cepstral highlights. We prepare an ANN with one single concealed layer, alluded to as ANN. The contribution to the ANNs are MFCC highlights with a few casings of going before and following setting. We don't pre-prepare the system.

3.5. Networks hyper-parameters

The hyper-parameters of the system are: the info window measure win, comparing to the setting brought with every model, the part width of the primary convolution, communicated in tests, the portion width kWn, the move dWn and the quantity of channels dn of the other nth convolution layers, the pooling width kWmp of max pooling layers and the shrouded layer width. They are tuned by early-halting on the approval set. The design is made out of 3 convolutional and max pooling layers and 1 concealed layer. The best execution on TIMIT was found with: 50 tests part width for the principal convolution, 310 ms of setting, 5 outlines piece width, 80, 60 and 60 channels, 500 concealed units and 3 pooling width. The ANN

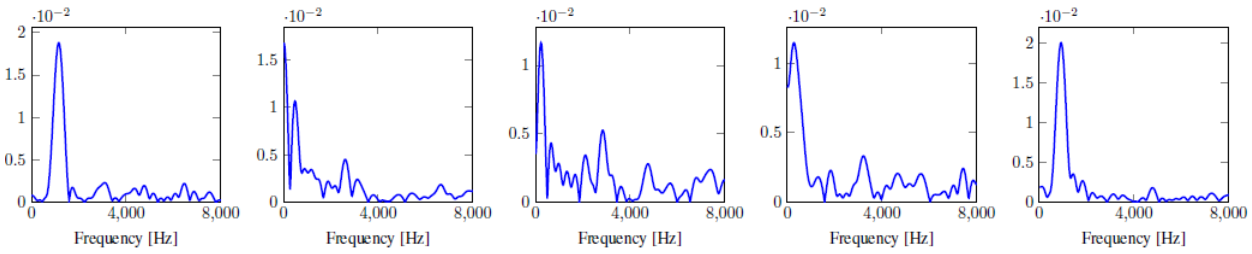

Figure 1: Illustration of the five most firing filters, with their proportion factor, for the center frame of phoneme /iy/.

standard uses 500 hubs for the shrouded layer. On Aurora2, 10-overlay cross-approval was utilized for tuning the hyper-parameters. The best execution was found with: 50 tests piece width

for the main convolution, 310 ms of setting, 7 outlines part width, 80, 60 and 60 channels, 500 concealed units and 3 pooling width. The ANN gauge utilizes 500 hubs for the concealed layer. The examinations were executed utilizing the torch7 tool kit [26].

4. FILTERS ANALYSIS

In the vast majority of the ongoing convolutional neural systems based frameworks proposed in the writing, the info highlights are either customary cepstral-based highlights [5] or otherworldly based portrayals, for example, Mel filter bank coefficients [10]. These highlights are typically processed on a 25 ms window, with a move of 10 ms. The key contrast in our framework is that the CNN takes the transient crude discourse straightforwardly as info. In this way, the primary convolutional layer should go about as a filter bank, learned in an information driven way. The part width of this first convolutional layer, speaking to the length of the worldly information window, was chosen exactly, on the approval set. The best execution was found with a short window, around 2-4 ms discourse, with a move of 0.6 ms, or, in other words shorter than in regular cepstral-based highlights preparing. The channels learned by the primary convolution can be viewed as coordinating channels. In our past investigations, we demonstrated that they react to different occurrence transfer speeds [16], and that they demonstrate some level of invariance crosswise over databases [17].

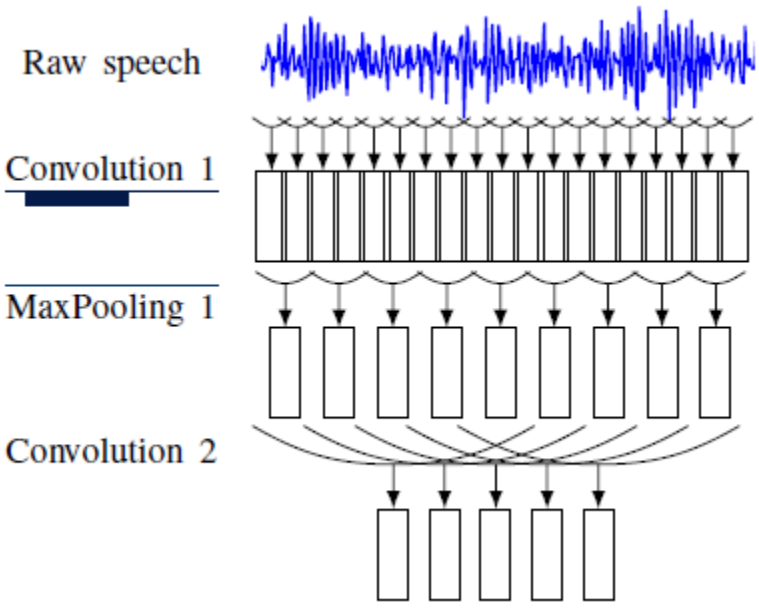

Figure 2: Detailed view of the first two stages of the CNN. The outputs of the first convolution are combined in the second convolution.

As outlined in Figure 2, in the proposed design the yields of the primary convolution layer which comprises of a bank of coordinated channels are straightly joined subsequent to maxpooling task. We take a disentangled perspective of this procedure where a discourse flag relating to a sound is gone through a bank of direct time invariant channels and the yields are then joined straightly. In such a case, we can imagine the aggregate recurrence reaction of the channel banks after straight blend as an entirety of the recurrence reactions of the channels that are terminating/coordinating to the unearthly qualities of the sound. Utilizing this disentangled view, we considered the normal ghostly attributes of the vowel sound that is being displayed in the accompanying way: The inside casing for a given vowel in a succession is chosen and sent to the system.

• We figure out which channels are terminating the most for each casing by taking the argmax.

• The initial two tasks are rehashed over the entire approval informational index monitoring the

number of times ni the channel I is activated.

• The recurrence reaction Fi of each channel I is figured by taking the size of the

Fourier Transform and normalizing it.

• Finally, the mean recurrence reaction for a given vowel f vowel is registered by including the recurrence reactions of the five most terminating channels, weighted by an extent factor I, or, in other words the quantity of time the channel I is activated, standardized by the aggregate number of appearances:

$$\lambda_i = \frac{n_i}{\sum_j n_j}.$$

$$f_{vowel} = \sum_i \lambda_i F_i$$

The quantity of terminating channels was set to five, since it speaks to the vast majority of the commitment to the channels yield. A representation of the five most terminating channels for one phoneme/iy/is given in Figure 1. The recurrence reactions of chose vowels, registered on the approval set of TIMIT are exhibited in Figure 3. The swells present in the plots can be credited to the way that these channels are found out on sub-segmental discourse flag, i.e. a length of 2-4 ms. It could be seen that the normal recurrence reaction resembles a smooth otherworldly envelope. Given that we could estimate that in the confounded (loud) conditions the CNN based framework ought to have a pattern like standard cepstral highlights, which tends to display unearthly envelope data (of around 25ms discourse flag). We learn this viewpoint in the

following segment. Table 1: Results on the Aurora tests A and B, given in Word Recognition Rate (WRR), arrived at the midpoint of over the four clamors. The two frameworks have 250k parameters.

	Test A							Test B						
SNR [dB]	clean	20	15	10	5	0	-5	clean	20	15	10	5	0	-5
	Clean Training													
ANN	96.9	86.1	74.1	51.4	25.5	13.9	10.1	97.4	87.3	77.8	59.8	33.5	15.0	8.9
CNN	97.3	88.3	76.1	53.0	24.7	11.2	8.0	97.2	90.4	83.2	64.9	38.7	19.1	10.1
	Multi-conditional Training													
ANN	92.1	91.6	89.0	83.4	70.0	38.2	14.5	92.1	85.1	80.9	73.8	59.7	34.1	14.5
CNN	97.6	97.4	96.6	93.9	84.8	55.1	19.5	97.6	94.8	93.4	89.0	77.4	48.0	18.7

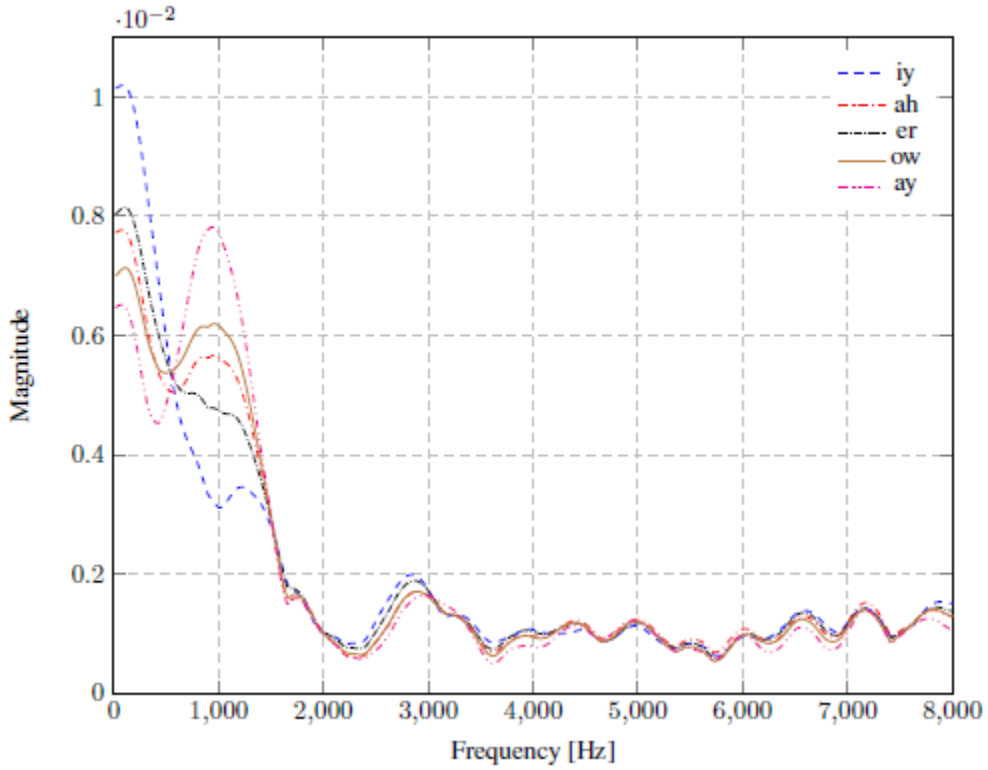

Figure 3: Mean frequency responses on the TIMIT validation set for phonemes /iy/;/ah/;/er/; /ow/ and /ay/.

5. NOISE STUDIES

DNN and CNN based frameworks had been appeared to yield state-of-the-workmanship results in ASR. Be that as it may, this information driven methodology could bring up issues about the powerlessness of the framework to confounded conditions. In this area, we present clamor vigor ponders on two assignments: associated word acknowledgment errand on the Aurora2 corpus and phoneme acknowledgment undertaking on the TIMIT corpus. For a reasonable correlation, in these investigations we do not play out any further standardization on the MFCC highlights, for example, cepstral mean standardization. The reason being that such normalizations could be imagined at flag level as separating tasks. Concentrate in detail these perspectives is open for further research and as talked about later in Section 6, is a piece of future work.

5.1. Word recognition study

Table 1 introduces the outcomes on Aurora2 corpus. It tends to be seen that in both clean condition preparing and multi-condition preparing, the CNN-based framework beats the ANN-based framework. This can be seen noticeably on account of multi condition preparing. The execution of the CNN-based framework is like the main revealed framework on Aurora2 corpus [24]. The feeble execution of the ANN-based framework could be credited to absence of highlight standardization and low limit.

Table 2: Results for the clean and multi-condition training on the TIMIT core test set, given in PRR. [28]

SNR [dB]	ANN		CNN	
Training	clean	multi	clean	multi
30dB	52.5	54.3	65.5	66.8
25dB	46.7	50.8	59.7	64.8
20dB	40.3	46.6	50.5	60.8
15dB	32.7	41.1	39.1	53.5
10dB	26.1	34.2	27.8	42.8
5dB	21.2	26.4	18.3	30.8
0dB	17.4	20.2	9.9	21.4

5.2. Phoneme recognition study

Table 2 displays the outcomes on TIMIT corpus for the standard and the CNN-based framework, communicated in term of Phoneme Recognition Rate (PRR). On account of clean preparing, it very well may be seen that the CNN-based framework is somewhat more hearty than the

benchmark. Notwithstanding, the execution of the CNN-based framework debases at low SNR level contrasted with the standard. This could be because of the little measure of fluctuation accessible in the TIMIT corpus, prompting channels which don't sum up exceptionally well to jumbled conditions, as of now appeared in [17]. On account of multi-contingent preparing, it could be seen that the CNN-based framework is reliably more vigorous than the gauge. By and large, these outcomes show that the CNN based framework pursues a comparable pattern to standard framework utilizing cepstral-based highlights as information.

6. SUMMARY AND FUTURE WORK

In synopsis, the channel investigation think about shows that the highlights learned between the initial two convolution layers of the CNN tends to demonstrate the ghostly envelope of sub-segmental discourse flag. The commotion hearty ASR examines demonstrate that these highlights are helpless to clamor yet not to indistinguishable degree from MFCC highlights (with no standardization). To enhance the strength of the CNN-based framework, we can abuse the parallel between time space preparing and recurrence area handling. For example, we could enhance the power by sifting the discourse flag utilizing the Wiener channel method in the Aurora Advanced Front End [27] and afterward encouraging it into the CNN. Our future work will explore these perspectives and will ponder in examination with commotion vigorous ghostly based component extraction.

REFERENCES

[1] H. Bourlard and N. M, Connectionist communication detection: a hybrid approach. Springer, 1994, vol. 247.

[2] G. E. Hinton, S. Osindero, and Y. W. Teh, “A fast learning algorithm for deep belief nets,” Neural computation, vol. 18, no. 7, pp.1527–1554, 2006.

[3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly,A. Senior, V. Vanhoucke, P. Nguyen, and T. N. Sainath, “Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups,” SPM,IEEE, vol. 29, no. 6, p. 8297, 2012.

[4] F. Seide, G. Li, and D. Yu, “Conversational speech transcription using context-dependent deep neural networks,” in Procedure. of Interspeech,2011, pp. 437–440.

[5] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-depend entpre-trained deep neural networks for large-vocabulary speech recognition,” IEEE Transactions on Acoustic, Language, and Verbal Treating, vol. 20, no. 1, p. 3042, 2012.

[6] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 14 –22, jan. 2012.

[7] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in ANIPS22, 2009, pp. 1096–1104.

[8] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," in Proc. of ICASSP, 2012, pp.4277–4280.

[9] T. N. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran,"Deep convolutional neural networks for lvcsr," in Proc. OfICASSP, 2013, pp. 8614–8618.

[10] P. Swietojanski, A. Ghoshal, and S. Renals, "Convolutional neural networks for distant speech recognition," SPL,IEEE, vol. 21, no. 9, pp. 1120–1124, September 2014.

[11] E. Bocchieri and D. Dimitriadis, "Investigating deep neural network based transforms of robust audio features for lvcsr," in Proc.of ICASSP, 2013, pp. 6709–6713.

[12] N. Jaitly and G. Hinton, "Learning a better representation of speech sound waves using restricted boltzmann machines," inProc. of ICASSP, 2011, pp. 5884–5887.

[13] T. Sainath, B. Kingsbury, A.-R. Mohamed, and B. Ramabhadran,"Learning filter banks within a deep neural network framework,"in Proc. of ASRU, Dec. 2013, pp. 297–302.

[14] Z. T¨uske, P. Golik, R. Schl¨uter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for lvcsr," in Proc. of Interspeech, Singapore, Sep. 2014, pp. 890–894.

[15] D. Palaz, R. Collobert, and M. Magimai. -Doss, "End-to-end Phoneme Sequence Recognition using Convolutional Neural Networks,"ArXiv e-prints, Dec. 2013.

[16] D. Palaz, R. Collobert, and M. Magimai.-Doss, "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in Proc. of Interspeech,2013.

[17] D. Palaz, M. Magimai.-Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in Proc. of ICASSP, April 2015.

[18] Y. LeCun, "Generalization and network design strategies," in Connectionism in Perspective, R. Pfeifer, Z. Schreter, F. Fogelman,and L. Steels, Eds. Zurich, Switzerland: Elsevier, 1989.

[19] J. Bridle, "Probabilistic interpretation of feed forward classification network outputs, with relationships to statistical pattern recognition," in Neuro-computing: Algorithms, Architectures and Applications, 1990, pp. 227–236.

[20] L. Bottou, "Stochastic gradient learning in neural networks," in Proceedings of Neuro-Nmes 91. Nimes, France: EC2, 1991.

[21] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden markov models," IEEE Transactions on Acoustics,Language and Signal Processing, vol. 37, no. 11, pp. 1641–1648, 1989.

[22] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech communication, vol. 12, no. 3, pp. 247–251, 1993.

[23] H.-G. Hirsch, "Fant-filtering and noise adding tool," 2005.

[24] H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCATRW (ITRW), 2000.

[25] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason,V. Valtchev, and P. Woodland, "The htk book," Cambridge University Engineering Department, vol. 3, 2002.

[26] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: Amatlab-like environment for machine learning," in BigLearn,NIPS Workshop, 2011.

[27] H.-G. Hirsch and D. Pearce, "Applying the advanced ETSI frontend to the aurora-2 task," Tech. Rep., 2006, version 1.1.

[28] Dimitri Palaz1;2, Mathew Magimai.-Doss1, Ronan Collobert3;Dimitri Palaz1;2, Mathew Magimai.-Doss1, Ronan Collobert