MACHINE LEARNING FOR SPOKEN LANGUAGE TECHNOLOGY

Dragoş BURILEANU[1], Șerban MIHALACHE[1], Valentin ANDREI[1], Alexandru-Lucian GEORGESCU[1], Horia CUCU[1], Corneliu BURILEANU[1]

Abstract. *Spoken language technology is one of the domains in which, in our days, machine learning algorithms and especially neural networks are used. Some applications will pe presented in this paper: detecting overlapped speech on short time frames (till 25ms), emotion recognition from speech (including speech stress detection and deceptive speech detection) and the performances of the last large vocabulary continuous speech recognition systems for Romanian developed in the SpeeD Laboratory, from Research Institute "CAMPUS", University POLITEHNICA of Bucharest*

Keywords: machine learning, overlapped speech, speech emotion recognition, Romanian speech recognition.

DOI https://doi.org/10.56082/annalsarsciinfo.2021.1-2.25

1. Detecting overlapped speech on short timeframes using deep learning

In several presentations in the Information Science and Technology Section of the Academy of Romanian Scientists I pointed out some of the main research directions for the Speech and Dialogue ("SpeeD") team. Now I am able to give more details about some achievements in several arias of interest: detecting overlapped speech on short timeframes, emotions recognition from speech, new approaches to Romanian speech and speaker recognition. What do these seemingly very different areas have in common?

I am trying to demonstrate that the methods offered by machine learning could provide viable solutions for the most diverse applications. But it is also an opportunity to share some of the achievements of the team I am working with.

Long speech frames, i.e. more than 500 ms, have a higher probability of containing partially overlapped speech (e.g. one speaker produces an utterance 200 ms after another one has started). This leads to risk of decreasing accuracy for: blind speech source separation (BSS), speaker identification, crowd-sensing. Detecting overlapped speech on short timeframes can contribute to key BSS applications.

[1] Speech & Dialogue Laboratory (SpeeD), Research Institute „CAMPUS", University POLITEHNICA of Bucharest
http://speed.pub.ro/

In some previous papers [13], [14] we presented several methods for competing speaker counting and compared them with the human capabilities of counting the speakers in an overlapped speech recorded on a single channel.

The Figure 1 shows that if there are more than 4 simultaneous speakers in a single channel recording, human listeners have serious difficulties in counting them. Therefore, we limited our overlapped speech detection study to up to 3 simultaneous speakers.

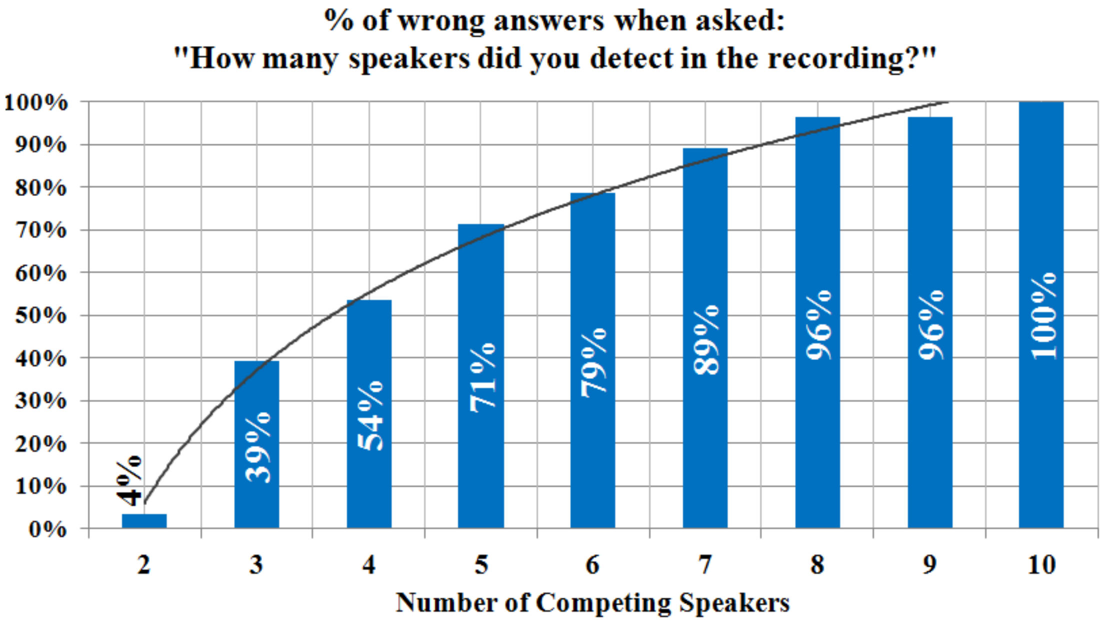


Figure 1. Number of competing speakers detected by human listeners

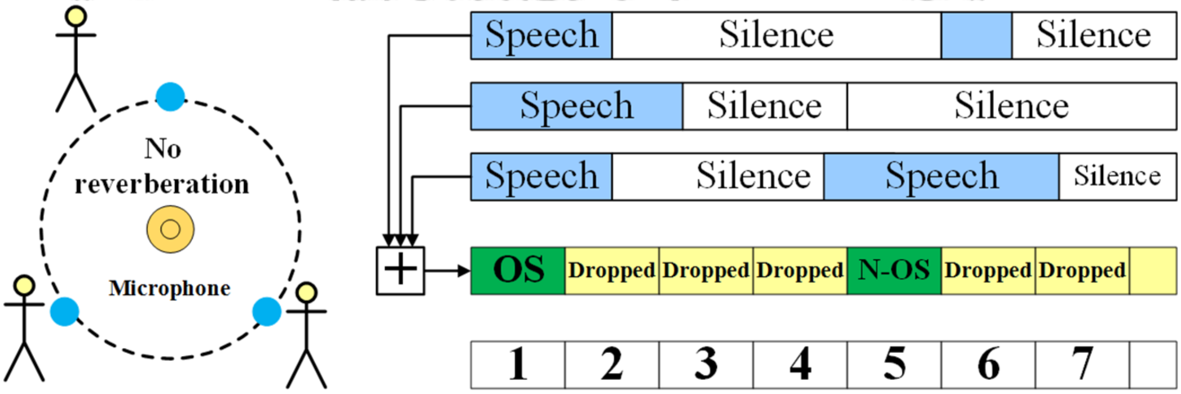


Figure 2. Speech source mixing

The setup of our experiments for training and inference will use up to 3 simultaneous speakers per mixture. The mixer normalizes the generated samples. 100k mixtures for training, 20k mixtures for inference. Speech source mixing needs to accurately label a timeframe. We only select timeframes with full overlapping (Figure 2).

1.1. Feature set selection

Providing unprocessed input to deep neural networks yields satisfactory results for image analysis applications, but for speech processing, feature engineering is still important.

So we used a set of extracted feature sets normalized, i.e. brought to similar numerical ranges to speed-up convergence>

- Signal's frequency spectrum. In our days this is the "raw / unprocessed" input. Contains the highest amount of information for the deep neural network to create features.
- MFCC coefficients as a dominant feature in prior work. We investigated the use of first and second order derived coefficients with no improvement in accuracy.
- Signal envelope computed with Hilbert Transform. An overlapped speech sample tends to have a flat shape and we expect the neural network to detect this.

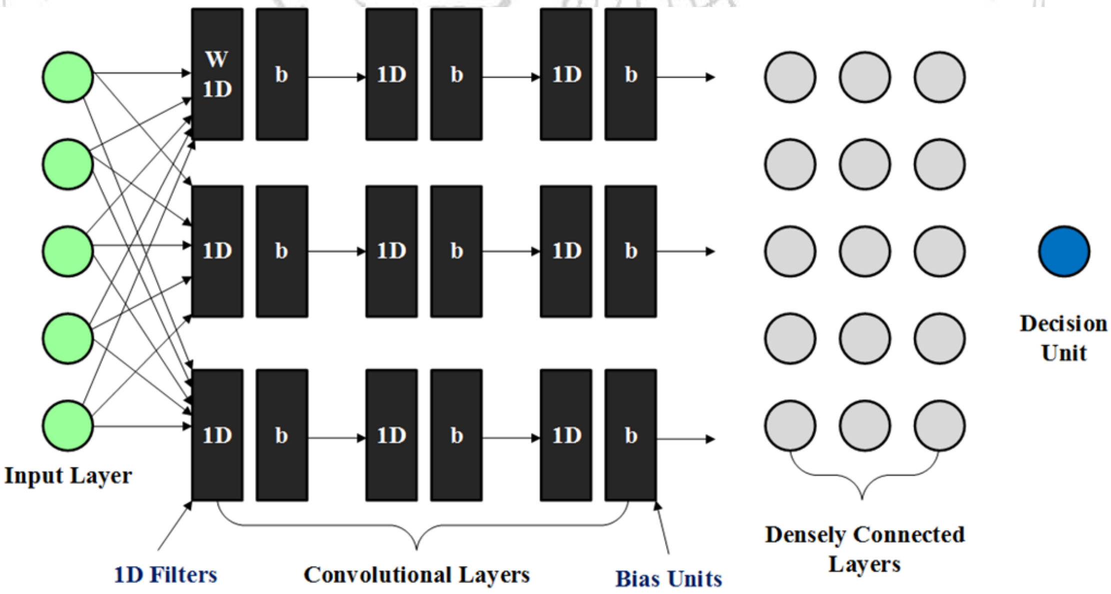


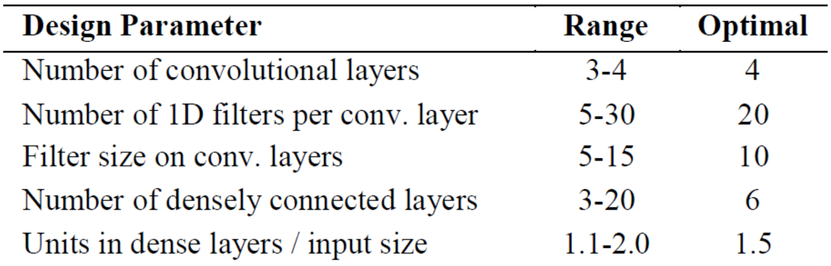
Figure 3. Deep Neural Network (DNN) architecture

- AR (autoregressive model) coefficients produces a wide numerical range that was observed to amplify subtle numerical differences between overlapped and non overlapped speech.
- We provide the option of adding squared features for all the components which can add extra performance when decision boundaries are very complex.

1.2. Deep Neural Network (DNN) architecture

We used convolutional layers as first layers due their accepted ability to create new feature sets [1], [2]. Then we considered that densely connected layers with convolutional layers are enough for the network to analyze the features (Figure 3).

Experiments were done in order to determine the optimal combination of DNN parameters (Table 1).

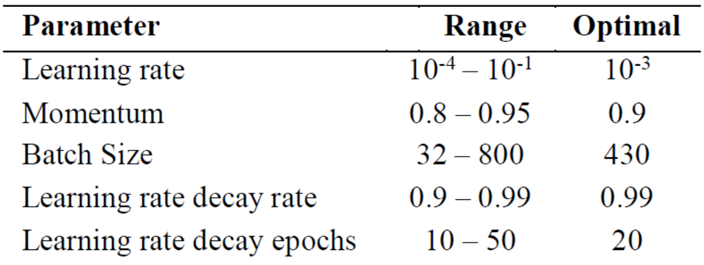


Design Parameter	Range	Optimal
Number of convolutional layers	3-4	4
Number of 1D filters per conv. layer	5-30	20
Filter size on conv. layers	5-15	10
Number of densely connected layers	3-20	6
Units in dense layers / input size	1.1-2.0	1.5

Table 1. DNN parameters

For the DNN training we used TensorFlow to create the entire experiments' infrastructure. Stochastic Gradient Descent was selected as the model weights update method by using the Momentum optimizer implemented in TensorFlow and activating the Nesterov Accelerated Gradient.

For the hyperparameters we considered some usual ranges as shown in Table 2.



Parameter	Range	Optimal
Learning rate	$10^{-4} – 10^{-1}$	10^{-3}
Momentum	0.8 – 0.95	0.9
Batch Size	32 – 800	430
Learning rate decay rate	0.9 – 0.99	0.99
Learning rate decay epochs	10 – 50	20

Table 2. DNN hyperparameters

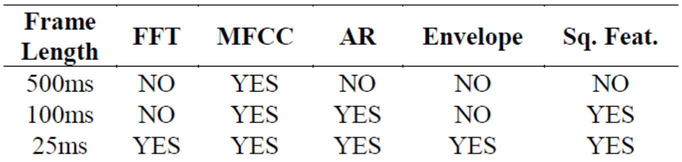
The learning rate was decayed across epochs to ensure convergence towards the end of the training, when the weights need to be updated in small steps. Surprising enough the convergence was achieved with such reduced values for learning rate. Batch size depends on multiple factors, but the most important is related to memory usage. Batch size and the other hyperparameters are intercorrelated.

The DNN architecture parameter ranges were intuitively selected based on the size of state of art models used in speech analysis (e.g. Deep Speech and Deep Speech 2). The optimal architecture for our dataset was determined after experimenting multiple combinations.

Hyperparameter ranges were identified after several sampling steps and some conclusions can be summarized: learning rate was the key tracked parameter; batch size was limited by the memory capacity of the system; the rest of parameters were intuitively selected based on widely known architectures

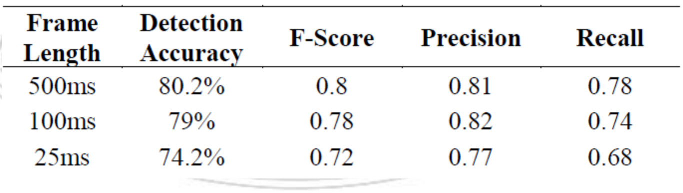
1.3. Results

We analyzed how the frame length affects the accuracy of the overlapped speech detection (remember, the frame length is important for adoption in various applications: e.g. longer frame lengths may be suitable for crowd sensing while short frame lengths can help BSS). The type selected features is presented in Table 3 and the detection performances using various measurements are presented in Table 4.



Frame Length	FFT	MFCC	AR	Envelope	Sq. Feat.
500ms	NO	YES	NO	NO	NO
100ms	NO	YES	YES	NO	YES
25ms	YES	YES	YES	YES	YES

Table 3. Feature selection per targeted case



Frame Length	Detection Accuracy	F-Score	Precision	Recall
500ms	80.2%	0.8	0.81	0.78
100ms	79%	0.78	0.82	0.74
25ms	74.2%	0.72	0.77	0.68

Table 4. Detection performance

Severeal conclusions can be summarised from the anlyze of the results:

- Longer frame lengths show improved accuracy because they contain much more information.

- For 500 and 100 ms frame durations, we could not use raw inputs (e.g. frequency spectrum) because the training duration grows exponentially.
- In order to obtain reasonable accuracy for short time frames we needed to add features.
- We speculate that additional features may improve the F-Score of the detection.
- Overlapped speech detection can be achieved successfully with the proposed DNN architecture composed out of convolutional layers and multi-layer-perceptron.
- We obtained an F-Score of 0.72 for 25 ms timeframes. In existing literature, the highest F-Score – collected in circumstances similar to our experiment – was reported with a value of 0.63.
- Longer frames may yield better accuracy, though longer frames my limit the applications where this method can be adopted – crowd-sensing is one of them.

2. Emotion Recognition from Speech

The broad framework of this topic is the dissimulated behavior monitoring. We considered the following tasks: Speech Emotion Recognition (SER), Speech Stress Detection (SSD) and Deceptive Speech Detection (DSD). The target applications are in forensics (questionings, interviews etc.), surveillance (suspicious behavior), medical (monitoring / prevention – stress, anxiety, depression) etc.

2.1. Classifications and main characteristics

The important features of every task are:

For Speech Emotion Recognition (SER):

a) Discrete categories (classification):

- each emotion is a separate class, with its own characteristics;
- typical system: 4-7 emotional classes;
- additional interest: just negative emotions (reduced set); any emotion vs. neutral (binary); each emotion vs. its absence (binary).

b) Dimensional models (continuous; regression – Figure 4):

- several properties (dimensions) determine an “affective space” in which each emotion class is defined by a certain sub-space;
- typical system: 2D space (plane) (arousal & valence);
- two possible annotations: a single value pair for every file in a database (*global annotation*) OR quasi-continuous annotation for every short time frame (*sequential annotation*).

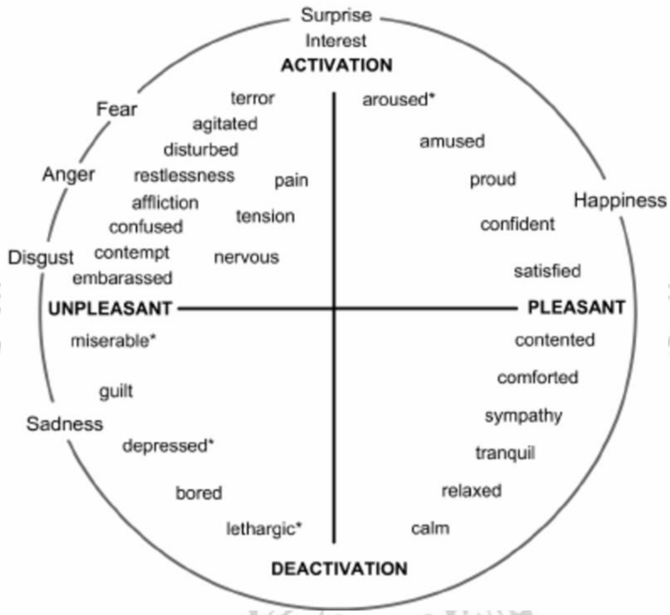


Figure 4. Dimensional model for SER

For Speech Stress Detection (SSD):

Discrete categories (classification):

- various types of stress (associated with speaking style, ambient conditions, problem solving etc.);
- typical system: 3-6 classes (out of a pool of 11-16);
- additional interest: any type vs. neutral (binary); each type vs. its absence (binary);
- additional approach (indirect classification): using a proxy for stress, e.g. fear (due to scarcity of available direct data).

For Deceptive Speech Detection (DSD):

Discrete categories (classification):

- simple binary classification: deceptive (untruthful) vs. non-deceptive (truthful) speech;
- main interest: global untruthfulness (annotations available for subjects lying at the utterance / phrase / turn level, even if parts of their speech are truthful);
- alternative: local untruthfulness (annotations available for subjects lying at the segment / short phrase level).

2.2. Databases

Building an annotated database for SER, SSD or DSD is a very difficult task taking into account some particularities:

- simulated data, using actors or amateur speakers (e.g. students) are often of little relevance for real-life situations;
- it is very difficult to control the environment / scenario because predictable could mean lack of spontaneity;
- subjective data annotation – is the most important drawback. Because of this poor objectivity we shall deal with relatively unreliable data.

So, we investigated the following opensource data bases:

a) SER – discrete categories:

- EMODB (Berlin Database of Emotional Speech): German; 10 speakers; 535 recs.; 7 classes (e.g. Anger, Disgust, Fear, Sadness etc.);
- IEMOCAP (Interactive Emotional Dyadic Motion Capture Database): English; 10 speakers; 10039 recs.; 6 usable classes (e.g. Anger, Sadness etc.);
- CREMAD (Crowd-sourced Emotional Multimodal Actors Dataset): English; 91 speakers; 7442 recs.; 6 classes (e.g. Anger, Disgust, Fear, Sadness etc.).

b) SER – dimensional models:

- IEMOCAP (“IEMOCAP 2”): secondary annotation for IEMOCAP; 2 dimensions (arousal, valence, dominance)
- RECOLA (Remote Collaborative and Affective Interaction): French; 46 speakers; 23 long recs. (~3500 utterances); 2 dimensions (arousal, valence)

SSD:

- SUSAS (Speech Under Simulated and Actual Stress): English; 16 speakers; 14600 short recordings (35 keywords); 16 possible classes (various subsets of interest)
- SAFE (Situation Analysis in a Fictional and Emotional Corpus): English/French; 400 speakers; 400 medium recs. (~12000 utterances); 2 classes of interest (Fear and Neutral)

DSD:

- RLDD (Real-Life Trial Data for Deception Detection): English; 56 speakers; 121 medium recs. (~1800 utterances)
- **RODeCAR (Romanian Deva Criminal Investigation Audio Recordings)**: Romanian; 19 speakers; 25 very long recs. (~13500 utterances); authentic and reliable data (non-subjective annotation, high-stakes contexts etc.). **This is the most important resource because is annotated by our team on a reliable, genuine utterances obtained in real life in criminal investigations**.

We used featured hand-crafted or automatically extracted like acoustic and prosodic features: pitch, jitter, shimmer, Mel-frequency cepstral coefficients MFCC), loudness, low/high frequency spectral concentration etc. and additional modulation-based features derived from the instantaneous amplitude and frequency (speech as a series of AM-FM micro-modulations)

2.3. Deep Learning Models

We investigated several type of neural networks in order to find the best solution for our targets:

a) Multilayer perceptron (MLPs): used for the classification tasks (SER, SSD, DSD) and the global regression task (SER with arousal-valence utterance-level annotation):

- Standalone: offers good results (similar or better than the current state of the art).
- Ensemble classification: inspired by Support Vector Machine (SVM) multiclass approaches: One-vs-One (OvO) and One-vs-Rest (OvR).

b) Bidirectional recurrent neural networks (RNNs) with long short-term memory (LSTM) cells (Figure 5):

- used for all tasks, including the sequential regression task (SER with arousal-valence quasi-continuous annotation);
- used standalone.

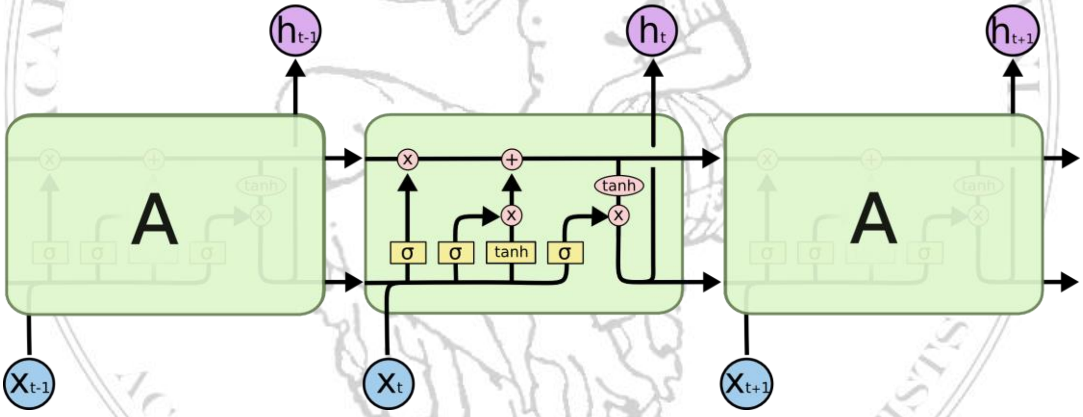


Figure 5. Bidirectional RNNs with long LSTM cells

c) Stacked autoencoders (SAEs- Figure 6):

- used for the classification tasks (SER, SSD, DSD) and the global regression task (SER with arousal-valence utterance-level annotations);
- used standalone;
- using MLPs as autoencoders.

2.4. Experimental Results

Several performance metrics were used in order to compare the performances of the SER systems:

- WA = weighted accuracy (ratio of all correct predictions VS. total no. of samples).
- UA = unweighted accuracy (average class recall, i.e. mean of per-class accuracies; more relevant for unbalanced classes).
- PCC = Pearson correlation coefficient (linear correlation).
- CCC = concordance correlation coefficient (PCC adjusted to also consider the prediction bias; more relevant for potentially mean-shifted data).

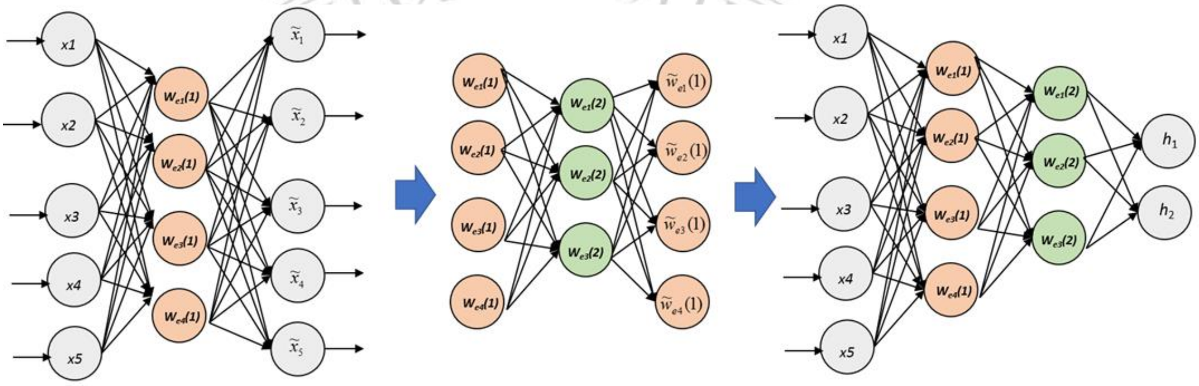


Figure 6. Stacked autoencoders

Table 5 presents the performances reported in some articles (“**Article**”- as they appear in the “References” section) compared with the results obtained by our team (**SpeeD**: Speech and Dialogue Research Laboratory), for different databases. The acronyms for the emotion states are:

ANG: Anger
HAP: Happy
EXC: Excited
SAD: Sadness
NEU: Neutral
FEA: Fear
FRU: Frustrated

“**Nclass**” stands for number of classes to discriminate among emotions and the results are evaluated in terms of different performance metrics presented above.

2.5. Conclusions

Current results are similar or better than the state of the art for SER with discrete categories on several databases and we are continuing to fine-tune the models in order to reach improved results on all considered datasets. We are still working to improve performance for SER with dimensional models (global annotation approach). The SSD and DSD tasks are still work in progress.

EMODB (Berlin Database of Emotional Speech):

Article	Nclass	Best Results
[28]	7 (all)	WA = 82.4%

[29]	4 (ANG, DIS, FEA, NEU)
	2 (EMO vs. NEU)
		WA = 84.3%
		WA = 94.9%

[30]	7 (all)	UA = 79.8%
[31]	7 (all)	WA = 81.5%

[33]	7 (all)
	5 (ANG, DIS, FEA, SAD, NEU)
		UA = 80.6%, WA = 81.6%
		UA = 87.3%, WA = 89.0%

SpeeD	7 (all)
	5 (ANG, DIS, FEA, SAD, NEU)
	2 (EMO vs. NEU)
		UA = 84.3%, **WA = 84.4%**
		UA = 93.2%, WA = 93.2%
		UA = 98.3%, WA = 98.2%

IEMOCAP (Interactive Emotional Dyadic Motion Capture Database):

Article	Nclass	Best Results
[22]	4 (ANG, HAP+EXC, SAD, NEU)	UA = 60.5%
[23]	4 (ANG+FRU, HAP+EXC, SAD, NEU)	UA = 48.7%, WA = 57.1% (Audio)
[27]	4 (ANG, HAP, SAD, NEU)	UA = 58.8%, WA = 63.5% (LLDs)
SpeeD	4 (ANG+FRU, HAP+EXC, SAD, NEU)	**UA = 58.7%, WA = 61.6%**

CREMAD (Crowd-sourced Emotional Multimodal Actors Dataset):

ART	Nclass	Best Results
[24]	6 (all)	WA = 57.2 % (Audio)
[25]	6 (all)	WA = 57.0% (Audio)
[26]	6 (all)	WA = 41.5% (Audio)

SpeeD	6 (all)
	5 (ANG, DIS, FEA, SAD, NEU)
		UA = 46.7%, WA = 51.6%
		UA = 50.8%, WA = 56.6%

IEMOCAP 2 (secondary annotation for IEMOCAP):

Article	Dimensions	Best Results

[32]	A-V-D (all)	PCC = 0.797 (Arousal)
		PCC = 0.566 (Valence)

SpeeD	A-V	**PCC = 0.679**, CCC = 0.637 (Arousal) **PCC = 0.382**, CCC = 0.321 (Valence)

Table 5. Performances of several SER systems

3. DNN Approach to Romanian Speech Recognition in "SpeeD" Laboratory

As I mentioned in some previous presentations in the Information Science and Technology Section of the Academy of Romanian Scientists, one of the most important activity of our team is to improve our results in the domain of spoken language technology. Some progress in Automatic Speech Recognition (ASR) systems will be presented in this paper [34], [35], [36].

3.1. Unsupervised Speech Corpus Extension

One assumption in our attempts is that different ASR systems make complementary errors. So, a possible method to improve performances of such a system is to transcribe unlabeled audio using two different ASR systems, align transcriptions and keep only identical parts and eventually use the timestamps provided by the ASR to cut the identical parts of the original audio file.

The details about an automatic annotation of unlabeled speech corpora and then how to improve the ASR systems by retraining using the new speech corpora are presented in the flowchart of the method (Figure 7).

There are plenty of characteristics to build the two complementary ASR systems if we consider the acoustic model type, the vocabulary size, the decoding language model complexity and/or the rescoring language model.

One important issue is to align and filter transcriptions. We used Dynamic Time Warping (DTW) to select common parts; then long sequences of consecutive word (if the number of characters exceeds a given threshold) are considered correctly transcribed and the time interval between two words must exclude the existence of intermediate un-transcribed words.

It is of a great importance to establish a correct method evaluation. The following performance figures are considered:

- Amount of speech selected after alignment relative to the total amount of unlabeled speech.
- Annotation quality for the selected speech measurable in word error rate (WER) and character error rate (ChER); can be computed using already annotated speech (reference) along with word level timestamps.

The speech corpora consist in:

- Read Speech Corpus (RSC): read and clean speech utterances in silent environment.
- Spontaneous Speech Corpus (SSC): spontaneous utterances from talk shows and news broadcasts.

- Unlabeled corpus: speech crawled from Romanian online media (two news websites and one radio station), during a 9 month period (Table 6).

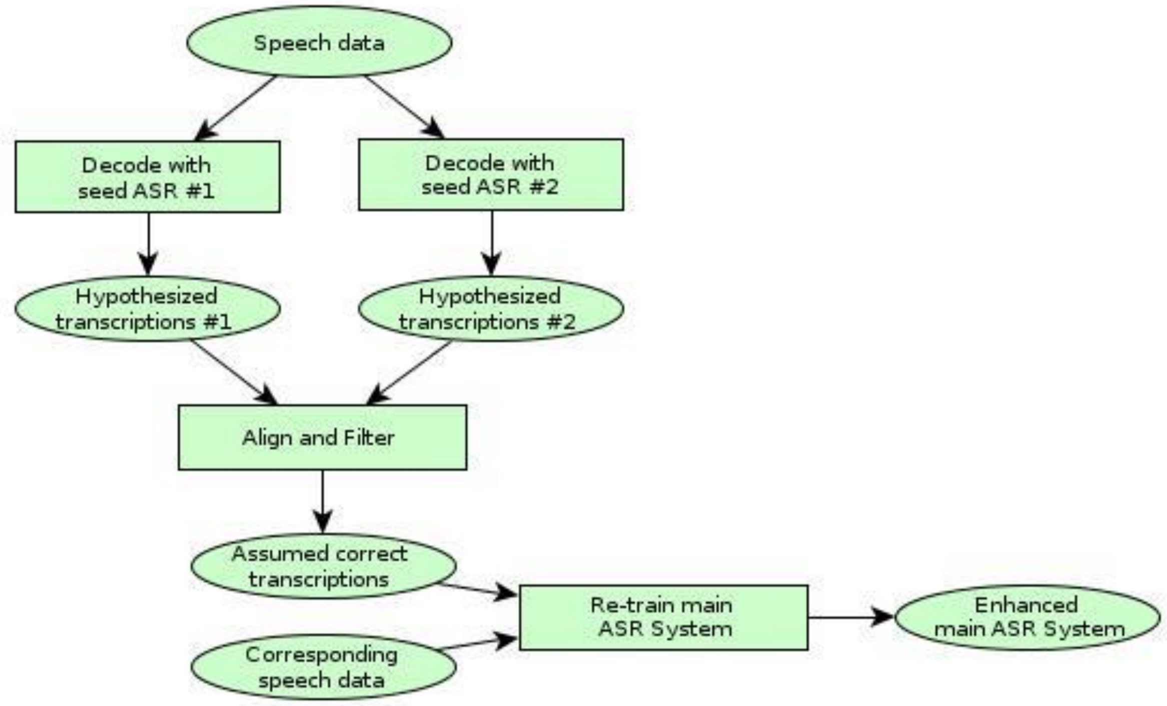


Figure 7. The flowchart of using two complementary ASR systems to improve overall performances

Purpose	Set	Size	
Training	RSC-train	94 h , 46 m	225 h, 31 m
	SSC-train	130 h, 44 m	
Evaluation	RSC-eval	5 h, 29 m	8 h, 58 m
	SSC-eval	3 h, 29 m	
Annotation	Source #1	367 h, 57 m	777 h, 53 m
	Source #2	331 h, 44 m	
	Source #3	78 h, 12 m	

Table 6. Speech corpora for unsupervised speech annotation experiments

The two ASR systems must be designed to be complementary, so they are different in many aspects:

ASR #1:

- The acoustic model is based on statistic models.
- The speech features are 13 mel-frequency cepstral coefficients (MFCC) plus their first and second order derivatives.
- The vocabulary is about 64k words.
- The language model (LM) is based on 3-gram statistics.
- No rescoring of the LM.

ASR #2:

- The acoustic model is based on time delay neural network (TDNN).
- The speech features are 40 MFCCs and 100-dimension iVectors.
- The vocabulary is 200k words.
- LM is based on 2-gram statistics.
- Rescoring for LM is using 4-gram statistics.

The results are not quite satisfactory: by doubling the amount of training speech data (adding 280 hours to the original 225 hours of speech), we obtained only 9% relative WER improvement. So, we concentrated on some other methods to improve the ASR systems accuracy.

3.2. ASR Improvements

In the last 5 years we targeted several directions to develop our large vocabulary continuous speech recognition (LVCSR) system:

- Speech and text resources acquisition.
- Improved language models: larger vocabulary, more grams for statical models.
- Improved acoustic models by switching from statistical models to deep neural network models.
- Speech feature transforms.
- Lattice rescoring after speech decoding.

The acoustic model is now based on Time Delay Neural Network (TDNN) which is able to learn long-term temporal dependencies. We are using as input 9 frames of relatively standard speech features: MFCCs and iVectors (especially useful for speaker adaptation). The input layer size is couple of thousand neurons and the output layer size is couple of hundred neurons. There are 3 - 6 hidden layers with around 1200 neurons. Framework and algorithms used are available in Kaldi ASR toolkit. Some experimental results are shown in Table 7.

TDNN Configuration	Number of training epochs	WER [%]	
		RSC-eval	SSC-eval
3500 in neurons 350 out neurons 6 hidden layers	1	6.4	21.7
	2	**6.2**	**21.0**
	3	6.3	20.7
	4	6.4	21.0
	5	6.4	21.2
	8	6.9	22.1

Table 7. Experimental results with TDNN used for the acoustic model

The Mel-frequency cepstral coefficients (MFCC) are extracted from 25 ms signal window length, shifted by 10 ms. The final feature vector has the dimension of 13 MFCCs x 9 frames. Supplementary, we used some features transforms:

- Cepstral mean and variance normalization (CMVN) in order to normalize the mean and variance of raw cepstra and eliminate inter-speaker and environment variations.
- Linear Discriminant Analysis (LDA) in order to reduce features space dimension keeping class discriminatory information.
- Maximum Linear Likelihood Transform (MLLT) to capture correlation between the feature vector components.

The improvement of the language model (LM) has the following characteristics:

- Kaldi ASR toolkit was used because it allows using LMs larger vocabularies (more than 64k words).
    - The text corpora used for language modeling was extended by collecting new texts from the Internet. We point out that text collected from the Internet needed diacritics restoration.
    - We have about 315M word tokens (in 2017)
    - Talk shows transcriptions (40M word tokens) already available.
- For the language models (LM):
    - Statistical n-gram models are used created by interpolating text corpora with various weights.
    - Various n-gram orders: from 1-gram to 5-gram.
    - Various vocabulary sizes: 64k, 100k, 150k and 200k words.

The results of language models' evaluations, without rescoring, are shown in Table 8. As expected, the best results are obtained for 200 kwords vocabulary with 3-gram models.

Vocabulary size	**ASR decoding LM order**	**WER [%]**	
		RSC-eval	**SSC-eval**
		w/o LM rescoring	
100 k words	1-gram	15.0	36.5
	2-gram	6.44	23.4
	3-gram	5.18	20.6
150 k words	1-gram	14.6	36.4
	2-gram	6.26	23.3
	3-gram	5.00	20.5
200 k words	1-gram	14.2	36.4
	2-gram	**5.90**	**23.2**
	3-gram	**4.62**	**20.5**

Table 8. LM evaluation without rescoring

When rescoring based on the algorithm presented above is used the results are slightly better as can be seen in Table 9.

Vocabulary size	**ASR decoding LM order**	**WER [%]**		**WER [%]**	
		RSC-eval	**SSC-eval**	**RSC-eval**	**SSC-eval**
		w/o LM rescoring		**with LM rescoring**	
100 k words	1-gram	15.0	36.5	6.06	22.5
	2-gram	6.44	23.4	5.04	20.3
	3-gram	5.18	20.6	5.05	20.1
150 k words	1-gram	14.6	36.4	5.81	22.4
	2-gram	6.26	23.3	4.85	20.3
	3-gram	5.00	20.5	4.85	20.1
200 k words	1-gram	14.2	36.4	5.39	22.4
	2-gram	**5.90**	**23.2**	**4.49**	**20.2**
	3-gram	**4.62**	**20.5**	**4.48**	**20.0**

Table 9. LM evaluation with rescoring

The overall improvement is shown in Table 10. Several conclusions can be pointed out:

- Several improvements of Speech and Dialogue Laboratory (SpeeD) LVCSR system for Romanian language were presented.
- The application of feature transforms, discriminative training and speaker adaptive training algorithms led to a lower WER.
- The use of DNN acoustic models is the most important change.
- Relative WER improvements between 20.7% to 30.8% over older statistic models.
- Increasing the LM size and the use of lattice rescoring triggered a lower WER.
- The overall relative WER improvement over the older system of about 5 years ago are: 70% on read speech and 48% on spontaneous speech.

SpeeD LVCSR System		**WER [%]**	
Acoustic model	**Language Model**	**RSC-eval**	**SSC-eval**
Statistic (CMU Sphinx, 2014)	64 k words, 3-gram	14.8	39.1
Statistic (CMU Sphinx, 2017)	64 k words, 3-gram	12.6	32.3
Statistic (Kaldi, 2017)	64 k words, 3-gram	9.0	26.4
DNN (Kaldi, 2017)	64 k words, 3-gram	6.2	21.0
	200 k words, 2-gram 4-gram (rescore)	**4.5**	**20.2**

Table 10. Overall improvements of our LVCSR system in the last 5 years

REFERENCES

[1] V. Andrei, H. Cucu, C. Burileanu, “Detecting overlapped speech on short timeframes using deep learning”, in the Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech), Stockholm, Sweden, 2017, pp. 1198-1202.

[2] Valentin Andrei, Horia Cucu, and Corneliu Burileanu, “Overlapped speech detection and competing speaker counting – humans vs. deep learning”, IEEE Journal of Selected Topics in Signal Processing, Vol. 13, Issue 4, Aug. 2019, pp. 850-862, Print ISSN: 1932-4553, Online ISSN: 1941-0484, DOI: 10.1109/JSTSP.2019.2910759, WOS:000477715300007

[3] K. Boakye, B. Trueba-Hornero, O. Vinyals, G. Friedland, “Overlapped speech detection for improved speaker diarization in multiparty meetings”, ICASSP 2008 – IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings, 2008

[4] W. Tsai, S. Liao, “Speaker Identification in Overlapped speech”, Journal of Information Science and Engineering, 2010, pp. 1891-1903

[5] R. Vipperla, J. T. Geiger, et. al., “Speech overlap detection and attribution using convolutive non-negative sparse coding”, ICASSP 2012 – Proceedings of International Conference on Acoustics, Speech and Signal Processing, 2012

[6] N. Shokouhi, A. Sathyanarayana, S. O. Sadjadi, J. H. L. Hansen, “Overlapped speech detection with applications to driver assessment for in-vehicle active safety systems”, ICASSP 2013 – Proceedings of International Conference on Acoustics, Speech and Signal Processing, 2013

[7] Garofolo, John, et al., “TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1”, Philadelphia: Linguistic Data Consortium, 1993.

[8] S. H. Yella, H. Bourlard, “Overlapped speech detection using long-term conversational features for speaker diarization in meeting room conversations”, IEEE/ACM Transactions on Audio, Speech and Language Processing, December 2014, Vol. 22, No. 12

[9] N. Shokouhi, A. Ziaei, A. Sangwan, J. H. L. Hansen, “Robust overlapped speech detection and its application in word-count estimation for prof-life-log data”, ICASSP 2015 – Proceedings of International Conference on Acoustics, Speech and Signal Processing, 2015

[10] S. A. Chowdhury, M. Danieli, G. Riccardi, “Annotating and categorizing competition in overlap speech”, ICASSP 2015 – Proceedings of International Conference on Acoustics, Speech and Signal Processing, 2015, pp. 5136-5121

[11] J. T. Geiger, F. Eyben, et. al., “Using linguistic information to detect overlapped speech”, INTERSPEECH 2013 – 15th Annual Conference of the International Speech Communication Association Proceedings, 2013

[12] J. T. Geiger, F. Eyben, B. Schuller, G. Rigoll, “Detecting overlapped speech with Long Short-Term Memory Recurrent Neural Networks”, INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association Proceedings, 2013

[13] V. Andrei, H. Cucu, A. Buzo, C. Burileanu, “Detecting the number of competing speakers – human selective hearing versus spectrogram distance based estimator,” INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association Proceedings, 2014, pp. 467 – 470

[14] V. Andrei, H. Cucu, A. Buzo, C. Burileanu, “Counting competing speakers in a timeframe – human versus computer”, INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association Proceedings, 2015, pp. 3999-4003

[15] V. Andrei, H. Cucu, A. Buzo, C. Burileanu, “Estimating competing speaker count for blind speech source separation”, SPED 2015 – Proceedings of 8th Conference on Speech Technology and Human Computer Dialogue, 2015, pp. 152–157

[16] J. Carletta, “Announcing the AMI Meeting Corpus”, The ELRA Newsletter 11(1), January–March 2006, p. 3-5

[17] Rainer Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics”, IEEE Trans. Speech and Audio Processing, July 2001, Vol. 9, pp. 504-512

[18] F. Grezes, J. Richards, A. Rosenberg, “Let Me Finish: Automatic Conflict Detection Using Speaker Overlap”, INTERSPEECH 2013 – 14th Annual Conference of the International Speech Communication Association Proceedings, 2013, pp. 200–204

[19] D. Amodei, S. Ananthanarayanan, et. al. “Deep Speech 2: End-to-End Speech Recognition in English and Mandarin”, Proceedings of the 33rd International Conference on Machine Learning, 2016, JMLR: W&CP Vol. 48

[20] Sutskever, J. Martens, G. Dahl, G. Hinton, “On the importance of initialization and momentum in deep learning”, ICML 2013 – Proceedings of International Conference on Machine Learning, 2013, pp. 1139-1147

[21] Lefter, I., Rothkrantz, L. J. M., & Burghouts, G. J. “A comparative study on automatic audio-visual fusion for aggression detection using meta-information”, Pattern Recognition Letters, 2013, Vol. 34(15), pp. 1953-1963.

[22] [1] S. Latif, M. Asim, R. Rana, S. Khalifa, R. Jurdak, and B. Schuller, “Augmenting GANs for SER,” in Proc. Interspeech 2020, pp. 521-525.

[23] [2] Z. Pan, Z. Luo, J. Yang, and H. Li, “Multi-Modal Attention for SER,” in Proc. Interspeech 2020, pp. 364-368.

[24] [3] E. Ghaleb, M. Popa, and S. Asteriadis, “Metric Learning-Based Multimodal Audio-Visual Emotion Recognition,” in IEEE Multimedia, vol. 27, iss. 1, pp. 37-48, Mar. 2020.

[25] [4] E. Ghaleb, M. Popa, and S. Asteriadis, “Multimodal and Temporal Perception of Audio-visual Cues for Emotion Recognition,” in Proc. 8th Intl. Conf. on Affective Computing and Intelligent Interaction (ACII), Cambridge, United Kingdom, pp. 552-558, Sep. 2019.

[26] [5] R. Beard et.al., “Multi-modal Sequence Fusion via Recursive Attention for Emotion Recognition,” in Proc. 22nd Conf. on Computational Natural Language Learning (CoNLL), Brussels, Belgium, pp. 251-259, Nov. 2018.

[27] [6] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic SER using RNNs with local attention,” in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, Louisiana, United States of America, pp. 2227-2231, Mar. 2017.

[28] [7] R. Lotfidereshgi and P. Gournay, “Biologically inspired SER,” in Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, Louisiana, United States of America, pp. 5135-5139, Mar. 2017.

[29] [8] J.C. Vasquez-Correa et.al., “Emotion recognition from speech under environmental noise conditions using wavelet decomposition,” in Proc. Intl. Carnahan Conference on Security Technology (ICCST), Taipei, Taiwan, pp. 247-252, Sep. 2015.

[30] [9] T. Chaspari, D. Dimitriadis, and P. Maragos, “Emotion classification of speech using modulation features,” in Proc. 22nd European Signal Processing Conf. (EUSIPCO), Lisbon, Portugal, pp. 1552-1556, Sep. 2014.

[31] [10] I. Bisio, A. Delfino, F. Lavagetto, M. Marchese, and A. Sciarrone, “Gender-Driven Emotion Recognition Through Speech Signals For Ambient Intelligence Applications,” in IEEE Trans. on Emerging Topics in Computing, vol. 1, iss. 2, pp. 244-257, Dec. 2013.

[32] [11] I. Guoth, M. Chmulik, J. Polacky, and M. Kuba, “Two-dimensional cepstrum analysis approach in emotion recognition from speech,” in Proc. 39th Intl. Conf. on Telecommunications and Signal Processing (TSP), Vienna, Austria, pp. 335-339, Jun. 2016.

[33] [12] S. Mihalache, D. Burileanu, G. Pop, and C. Burileanu, “Modulation-based SER with Reconstruction Error Feature Expansion,” in Proc. Intl. Conf. on Speech Technology and Human-Computer Dialogue (SpeD), Timișoara, Romania, pp. 1-6, Oct. 2019.

[34] A.-L. Georgescu, H. Cucu, C. Burileanu, “SpeeD’s DNN Approach to Romanian Speech Recognition,” in the Proceedings of the 9th Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, 2017, 8p, ISBN 978-1-5090-6496-0.

[35] A.-L. Georgescu, A. Caranica, H. Cucu, and C. Burileanu, “RoDigits – A Romanian Connected-Digits Speech Corpus for Automatic Speech and Speaker Recognition”, in UPB Scientific Bulletin Series C – Electrical Engineering and Computer Science, Ed. Politehnica Press, Bucharest, Vol. 80, Issue 3, pp. 45-62, 2018, ISSN: 2286-3540, WOS:000440896700004.

[36] Caranica, H. Cucu, A. Buzo, C. Burileanu, “On the Design of an Automatic Speaker Independent Digits Recognition System for Romanian Language”, in Journal of Control Engineering and Applied Informatics, vol. 18, no. 2, pp. 65-76, Jun 2016, ISSN 1454-8658, ISI IF 0.449.

.