LOST IN SEGMENTATION: THREE APPROACHES FOR SPEECH/NON-SPEECH DETECTION IN CONSUMER-PRODUCED VIDEOS

Benjamin Elizalde, Gerald Friedland

International Computer Science Institute, 1947 Center Street, Berkeley, CA 94704, USA
benmael@icsi.berkeley.edu, fractor@icsi.berkeley.edu

ABSTRACT

Traditional speech/non-speech segmentation systems have been designed for specific acoustic conditions, such as broadcast news or meetings. However, little research has been done on consumer-produced audio. This type of media is constantly growing and has complex characteristics such as low quality recordings, environmental noise and overlapping sounds. This paper discusses an evaluation of three different approaches for speech/non-speech detection on consumer-produced audio. The approaches are state-of-the-art speech/non-speech detectors–one based on Gaussian Mixture Models (GMM), another on Support Vector Machines (SVM), and the last on Neural Networks (NN). Using the TRECVID MED 2012 database, we designed training/testing sets combinations to aid the understanding of what speech/non-speech detection on consumer-produced media entails and how traditional approaches to this detection performed in this domain. The results revealed that the cross-domain state-of-the-art GMM and SVM systems' tests underperformed a one-layer NN algorithm, which had 20 % higher accuracy and computed audio 5 times faster.

Index Terms— audio segmentation, user-generated content neural networks, svm, gmm, speech non-speech

1. INTRODUCTION

Speech/non-speech detection is often seen as a developed field in speech processing, where only incremental improvement in specific domains seems to be possible. However, it is still a task with unresolved obstacles, especially when dealing with audio that does not adhere to a certain "domain," such as the audio in consumer-produced videos. Working with the audio in consumer-produced videos is important as these videos are the fastest-growing type of content on the Internet. YouTube alone claims that 72 hours of video are uploaded to its website alone every *minute*. These videos provide a wealth of audible information about the world. They consist of entertainment, instructions, personal records, and various aspects of life in general as it was when the video was recorded. Furthermore, there is information not only in the main focus of these videos, but also in the incidental and background audible context present in the videos. Each of these videos is a direct record of the world. As a collection, they represent a compendium of information that goes beyond what is captured in any individual recording. They provide information on trends, evidence of phenomena or events, social context, and societal dynamics. As a result, they are useful for qualitative and quantitative empirical research on a scale much larger than has ever been possible before. However, in order to make these videos accessible for research, we need to be able to automatically analyze the audible content of the recordings.

Historically, audio analysis research has been performed on corpora that were designed for a specific task, such as speech recognition, speaker identification, language recognition, etc. As a result, a seemingly simple task like speech/non-speech detection on consumer-produced videos raises completely new research questions. In general, the main problem of segmenting this audio resides in the combination of choosing a technical approach that will fit the audio best and dealing with the complexity and diversity of the audio audio. While different approaches for speech/non-speech detection exist, they have thus far only reached high accuracies for traditional, corpus-based, supervised segmentation tasks. The complication in segmenting "wild," consumer-produced audio is not only that one cannot rely on any single characteristic to draw boundaries between audio classes, but that it is difficult to pre-train models because of the high variance of the audio and the little availability of annotated audio sets.

The three approaches for speech/non-speech detection discussed in this paper present an evaluation of the problems associated with classifying this wild media. Each of the three evaluated approaches has been highly effective in processing traditional, corpus-based audio and are state-of-the-art speech/non-speech detectors. We used the TRECVID

We thank Adam Janin for his valuable technical support and discussion, and Nils Peters for his accurate advice. Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20066. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusion contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsement, either expressed or implied, of IARPA, DOI/NBC, or the U.S. Government.

MED 2012 to test the three approaches–one based on GMM, another on SVM, and the last on NN–by designing training/testing set combinations to aid the understanding of what speech/non-speech detection on consumer-produced media entails and how these traditional approaches to audio detection performed in this wild domain.

We structure the article as follows. Section 2 commences presenting some related work. Section 3 presents an overview of the three systems, while Section 4 describes the nature of the audio and the experimental setup. Section 5 submits the results and the corresponding analysis before Section 6 resumes with the conclusion and the outlook for the future.

2. RELATED WORK

Speech/non-speech segmentation has been well studied in controlled domains such as meetings and broadcast news, which involve the presence of background music and occasional noise, as summarized in [1].

A more complex acoustic domain currently under investigation is used in the DARPA RATS program [2]. This audio is collected under both controlled and uncontrolled field conditions over highly degraded, weak and/or noisy communication channels. For this scenario, [3] analyzes the performance of combining a one-layer NN and a GMM-based system along with an emphasis on feature extraction, which includes long span features and acoustic PLP features.

For consumer-produced data, research related to content analysis tasks includes speech activity detection (SAD), also known as voice activity detection (VAD), systems which use traditional hierarchical GMM based-algorithms. An example of the SAD task employing web videos [4] compared a GMM-based system using Mel Frequency Cepstral Coefficients (MFCCs) against a Maximum Entropy (MaxEnt) system using an alternative set of spectral and energy features. The article concluded that the MaxEnt and the set of alternative features yielded a lower error.

Despite this early promising conclusion that using a non-traditional method could bring better results, there seems to be no work that systematically evaluates different speech/non-speech classification algorithms on consumer-produced videos. We therefore concluded to present the beginning of such a study in this article.

3. SPEECH/NON-SPEECH DETECTION

The following section explains the three different approaches used in our comparison with a rationale as to why we chose each. Due to page limits constraints flow diagrams for each system weren't included. For further understanding refer to the corresponding publications.

Speech/non-speech segmentation systems have existed for several years. The most frequently used method is based on GMMs, which are trained on speech and non-speech audio segments, a Hidden Markov Model (HMM) as a temporal segmenter [1], and the Bayesian Information Criterion (BIC) [5] to compare models. GMMs have proven to be highly effective modeling distributions that exemplify an audio class such as speech and non-speech. It is also simple to create other audio models such as silence, to aid the segmentation. GMM techniques present drawbacks for large-scale data in that they are normally iterative and require significant computational power. Other weaknesses are that models may be less accurate if they aren't trained with enough data, and that they can be over fitted with the wrong number of mixtures.

In order to analyze the effectiveness of GMM systems on consumer-produced audio, we utilized the SHOUT [6] speech/non-speech system, which is considered state-of-the-art for meeting recognition [7]. It has two main characteristics. First, the system learns from the testing set instead of the training set. Second, this system does not have parameters to tune. This combination of characteristics makes SHOUT a self-sufficient, almost unsupervised algorithm and therefore an excellent option when little-to-no training data is available. During the training stage, two bootstrap GMMs must be first trained, one for speech and one for non-speech. In the testing stage, after the feature extraction of the audio, a bootstrap segmentation of speech and non-speech is performed. This segmentation is used to train silence and audible non-speech GMM models iteratively from the non-speech segments. After, a speech model is similarly trained from the speech segments. Once the three models are obtained, the necessity of the audible non-speech model is reviewed. This check is performed using the BIC, which is a penalized version of the maximum likelihood approach. The speech model and the audible non-speech model are then compared to see if they are the same. If they are, then the audible non-speech model is not needed and it is discarded while two new models are created speech and silence, if not then all of them are kept. This is performed iteratively to either merge the models or discard them. After merging the models, a retraining of the GMMs is performed.

Another technique for speech/non-speech detection is based on SVMs which classifies "bag of words" (here: frames) [8]. This approach has the potential to provide a solution to the wild-audio segmentation task, since SVMs have been designed to solve high-dimensional classification problems. Nevertheless, the use of SVMs for this task is not straightforward because temporality matters and SVM classifiers usually require an input of fixed length. Some of the disadvantages include that SVMs impose a binary classification and do not allow multiclass capabilities, although there are suggested solutions to these issues. Furthermore, SVMs are considered to be "shallow" architecture since they consist only on one-layer of a fixed kernel function, which needs to be carefully chosen and tuned according to the data type for best results.

For our study, we used a hybrid segmentation SVM sys-

tem [9] which consists of both unsupervised clustering and supervised classification. The system is divided in two stages of training and testing. For the training stage, there are two subsections. The first is an unsupervised section where a codebook is created with the output of a compilation of MFCCs that are fed into a K-means algorithm. The second is a supervised section in which the audio and the ground truth are used to create one set of labeled MFCC segments for speech and one set of labeled non-speech. The codebook and the sets of the labeled segments are then combined to create histograms. The number of histograms sets is determined by the number of sets of labeled segments-one set for speech and one for non-speech. These histogram sets are used to train a SVM model using a radial basis function kernel. For the testing stage, a set of unlabeled consecutive test segments for each audio file is created, which later is transformed into histograms in the same manner as the training stage. These histograms are classified by the SVM using the previously mentioned SVM model.

In contrast with the previous systems, the NNs are algorithms that may contain one or more hidden layers with parameterized non-linear modules that are subject to learning. NNs are hardly over-fitted and work well with multidimensional features. They also are quick and work well in multiclass tasks because they provide a probability space for data, facilitating the differentiation of samples. Furthermore, NNs provide an advantage over SVMs in that they can achieve similar performance with a smaller first layer, since the parameters of the first layer can be optimized for the task [10]. However, some of the disadvantages lay in the difficulty of training several layers and the possibility of the presence of local minima.

The NN based system for our study is a supervised approach [11] that consists of two parts of a NN with one hidden layer and a Finite State Transducer (FST). Note that in the paper an HMM Tandem method was used instead of the FST [12]. After the feature extraction of the training stage, a context window is created with a group of frames to train the NN, one state for speech and another for non-speech. Then the FST analyzes the statistics of the ground truth and creates a model containing transition probabilities and determining minimum duration values for speech and non-speech. In the testing stage, a context window of the testing file is input to the NN with the output being a likelihood score of the two classes for each frame. This score file is then passed to the FST along with the created model to determine the boundaries and duration for speech and non-speech segments. The FST ensures that the temporality of the segments is maintained. The NN system is the only one that included and ad-hoc FST, but it should not be the decisive factor in terms of segmentation performance.

4. DATA & EXPERIMENTAL SETUP

For the study presented herein, we chose to compare the performance of the three above described systems using a meeting dataset and the TRECVID 2012 consumer-produced video dataset.

4.1. Data

The ICSI Meeting Corpus is a collection of 74 meetings including simultaneous multi-channel audio recordings, word-level orthographic transcriptions, and supporting documentation collected at the International Computer Science Institute in Berkeley, ICSI. The meetings included are "natural" meetings in the sense that they would have occurred anyway. The meetings included here generally run just under an hour. This type of audio has little to no background noise or any other type of audible non-speech.

On the consumer-produced audio side, we used a subset of the NIST TRECVID MED 2012 video database called DEV-T. The entire dataset comprises a collection of training and testing data for a total of 150,000 video files of about three minutes each and with only 14 hours of manually annotated data by Language Data Consortium LDC. It is organized around 30 concept classes such as "Board Trick," "Feeding an Animal" or "Landing a Fish." In this type of data there is not a fixed structure. Music, unstructured speech, far field speech, high level background noise, and other examples of challenging to segment audible sounds may be encountered with some hard to identify and even unknown by human annotators.

Two randomly selected 12 hours subsets that included annotations for speech regions were used for the experiments. One subset corresponded to the TRECVID's DEV-T subsection and one to the Meeting corpora. Each corpus' training set consisted of 6 hours, as well as each corpus' testing set. In this article, meeting recordings will be referred to as "clean" data and the TRECVID 2012 as "wild" due to their above described characteristics. The ratio of speech and no-speech is 82% and 18% for clean and 55% and 45% for wild. From the wild data about 35% of the data include a type of overlap with other audio class such as music, singing, noise or unintelligible. All audio had a sample rate of a 16kHz, 1 channel, PCM format. The extracted audio features were typical 19 Mel Frequency Cepstral Coefficients plus Delta for a total of 38 dimensions, with a 25 ms window and a window step of 10 ms.

4.2. Experimental Setup and Error Metric

In order to normalize for the different characteristics of the three tested systems and to get an idea of their cross-domain adaptability, our experiments consisted of four training/testing combinations sets, one for each system for a total of 12 runs. The first combination set was clean/clean and was planned to test the systems with the best case scenario and

Set	Train/Test	GMM %	SVM %	NN %
1st	Clean/Clean	15	7	5
2nd	Clean/Wild	21	36	29
3rd	Wild/Clean	20	10	10
4th	Wild/Wild	26	37	19

Table 1. Speech/Non-Speech Error results for the 12 experimental runs.

obtain the performance baseline. The second and third sets included the combination of both datasets to observe how the systems behaved on mismatched conditions with consumer-data. The fourth and most important set, wild/wild, was intended to evaluate the best segmentation results in wild conditions. For each of the four experiments, a six hour subset of each database was used for the training stage and a different six hour subset for the testing stage.

In the fourth and last experiment two aspects of the processing time were compared–training and testing. Because the SVM and NN have a simple parallelization capability allowing the user to process several files from end to end at the same time, which the GMM system does not, only one file at a time was processed for each of the systems at the both stages. Therefore, none of the parallelization benefits were employed during either stage. The data used for the training stage was the same for the fourth experiment and the testing data was a one hour file from the meetings database. The three systems were run in a Dual Core AMD Opteron Processor 875 computer. The technical specs of this 64-bit machine included two CPU cores at a frequency of 2.2 GHz and 32 GB RAM.

The SAD systems were evaluated with the S_{error} % measurement [13]. This percentage relates the two speech error types: (MD) is the *Missed Detected* speech, or the total time of speech that was not classified as speech, and *False Alarm* (FA), which is the total time of non-speech that was falsely classified as speech. Lastly, S_{total} is the total time of speech in the ground truth. The equation writes as:

$$SAD_{error}\% = (MD + FA)/S_{total} * 100 \qquad (1)$$

While there are different distributions of speech and non-speech in the two datasets, the error metric converges at 45 % for the wild audio and 18 % for the clean audio, which is the expected random guess. These two error values could be use as a naive approach too, when assuming that every segment is speech.

5. RESULTS AND DISCUSSION

We commence our discussion of the results by presenting the raw numbers. The results on the performance of the four sets of experiments are shown in Table 1. The results on the computing time for training and testing are shown in Figure 1.

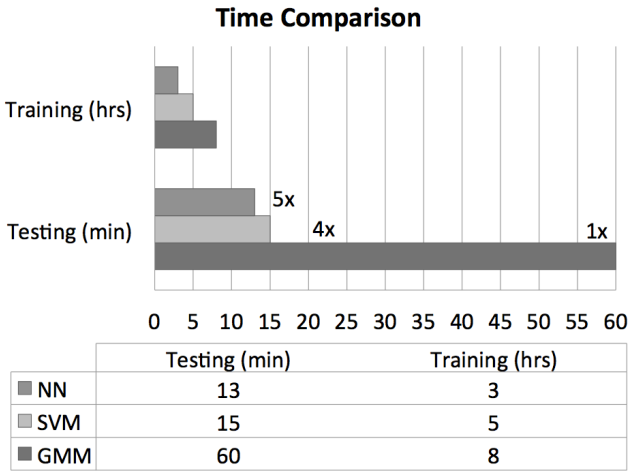


Fig. 1. Bottom-Line time requirements for the experiments.

This figure presents the time between processing the input audio into a segmentation output for the testing stage. For the training stage, the processing time between the input set of audio and the corresponding training model for each system is considered. Note that the SVM system's training time does not include the codebook creation. Lastly the Figure 2 contains the results of the FA and MD errors for each set of experiments and each of the three systems.

As predicted, the peak accuracy performance is observed in the clean/clean set. These values correspond to the best case scenario and represent the baseline performance. The NN approach outperforms the other approaches with 2 % less error than the SVM and 10 % less error than the GMM approach.

In the second set of experiments our working hypothesis (and literature experience) is confirmed in that all systems showed a significant decrease in the performance due to the mismatched conditions in classifying wild data while trained on clean audio. We interpret the fact that our tested GMM system performed with the lowest error rate as an indicator of quality for this particular system.

The third set of experiments serves to compare that mismatched conditions do not have such an impact on accuracy when classifying clean data trained on wild data. We assume that wild data is more representative of the clean case than the clean data is of the wild audio. This has been often assumed in literature and is one of the principal assumptions in the "Big Data" movement: With enough data, machine learning algorithms perform better in general [14]. Our results indicate that this is indeed true. The overall performance of all three systems is better here than in the second set. The SVM and the NN approaches had 10 % less error than the GMM counterpart which we interpret as a limitation of the GMM approach for "wild" audio classification.

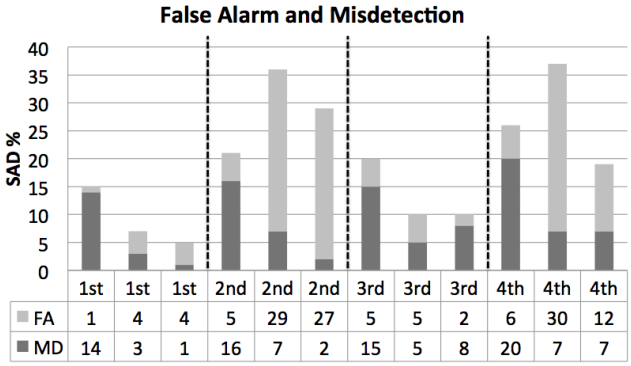
The mismatch conditions from the second and third experiments are not a phenomenon in consumer-produced audio

only. The two experiments results show how much the detection worsens in respect to the first experiment. The results also show how any combination of the mismatch conditions was not worse than the match conditions of the fourth experiment.

However, the speech from wild audio seems to be hardly representative of speech, due to the inherent characteristics such as environmental noises or overlap audio. Thus, confirming the principle assumption of this article for a need of cross-domain research in speech/non-speech detection. The wild data is the common factor on the difference in performance through the four experiments from each system. The last set of experiments, training on wild and testing on wild, lead to the lowest performance for the three systems. Surprisingly, for a one-layer NN approach, it yielded the best performance with 19 % error, with the GMM as its closest competitor. Therefore further study into how much improvement lays in multi-layered NN systems, such as Deep Belief Networks may be valuable.

The Figure 2 shows the two error types of the four sets of experiments for each of the three systems. In general, the GMM yielded a lower FA and the NN a lower MD. The comparison between the first set, which is the baseline, and the fourth set, which is the wild/wild set, returned an increment of both errors, specially the FA. The GMM system provides a lower FA through the four experiments in contrast to the NN. This is probably because GMM assumption of the speech distribution was better than the discriminative characteristics of the NN. The MD was high for the GMM, this might be because the speech GMM includes overlapped or environmental audio which cause confusion with some non-speech GMM. Important to mention is that singing occurs in the videos and is not labeled as speech, affecting the speech detection performance of the three systems.

Regarding computational demands, the GMM system was segmenting the audio in real time, which is at least 4 times slower than the other two systems, with the NN being the fastest. Each system had specific stages where audio was processed at a slower rate. The GMM algorithm was slow at iteratively retraining the GMMs after finding homogeneous models. Tuning will certainly make this algorithm faster, for instance reducing the number of components or increasing the size of the bootstrap segments will cut down in iterations. The SVM system contained no slow points during the testing stage. However, the main time concern of this algorithm, the creation of the codebook and the SVM model training, were not considered which would have provided a slow point. For the NN, the slow point resided in the FST step at the testing stage. In this step, the computation of the most adequate string line of probabilities to decide for speech and non-speech was a slow point and also consumed a significant amount of memory.



	1st	1st	1st	2nd	2nd	2nd	3rd	3rd	3rd	4th	4th	4th
FA	1	4	4	5	29	27	5	5	2	6	30	12
MD	14	3	1	16	7	2	15	5	8	20	7	7

Fig. 2. Two error types of the four sets of experiments for the GMM, SVM, and NN systems

5.1. Interpretation of the Results

We used a state-of-the-art GMM-based system that was optimized to deal with noise and mismatched conditions. As a result its performance in the clean/clean case was suboptimal. However, its performance degraded the least for the other three experimental sets, making a case for being a quite stable approach. As discussed previously, the major disadvantage of GMM segmentation is that the models need to be trained on a matching set. If the acoustic characteristics of the audio under evaluation are too different from the characteristics of the training set, the accuracy of the segmentation will be poor.

The SVM-based system performed comparably well to the NN system in clean audio. Despite what they were designed for and based on our results and the literature (see Section 2), SVM systems do not seem to cope well with the task of classifying wild audio. The advantage of using a codebook system for this study is that it allows us to analyze the generated code entries and their resulting use. It turns out that all codes ended up being very similar, but at the same time very few codebook entries were used with high frequency. This leads us to further confirm the conclusion that current SVM approaches seem not to be promising for this task. Note the consistency of the SVM system when testing the clean sets regardless of the training nature. This was because the codebook was compensating the variability of the training sets.

While NNs showed to be affected by the mismatch conditions, they definitely seemed to be outperforming both of the other systems in accuracy and computational performance. While we were surprised by the accuracy numbers, the reason for the computational performance efficiency is that NN approaches have been around for the longest time, leaving enough time for optimization. As result, an NN approach seems to be the best fit in terms of performance and time. We observed that the NNs are very good at classifying wild audio because they better allow for soft decisions. This is in contrast to the GMM approaches, which use a maximum

likelihood approach forcing their decision to select one of many. SVM approaches divide the hyper-plane forcing samples to belong to either one or the other, thus the division of the hyper-plane is not accurate when the sample types are not very distinguishable. The NN-based approach, however, outputs numerical vectors containing the probability of a frame belonging to either audio class. This allows the temporal segmenter to make a decision based on a "softer" margin.

All the systems were used off-the-shelf and improvements are expected on all of them if proper tuning is done. These systems were selected based on their performance on published literature and thought to have the highest likelihood of working with wild audio, but a study like this can always be arguable.

6. CONCLUSIONS AND FUTURE WORK

Speech/non-speech detection and audio segmentation in general are important front-end tasks of higher level audio processing. With wild audio becoming the main source of multimedia information from the internet, the importance of being able to accurately segment such cross-domain audio is more important than ever. Currently, little research has been done to analyze the audio segmentation task for consumer-produced media. With a completely generic, cross-domain learning algorithm being unavailable, this paper presented three different approaches to the task, which revealed that the NN-based system performed at least 20 % better and up to 5 times faster on consumer-produced audio than the other two approachesThe NN and the SVM approaches provided almost similar MD errors. The traditional GMM-based technique had a competitive performance especially in terms of FA. While the SVM system is almost as fast as the NN it showed low accuracy at classifying the audio tracks of consumer-produced videos. We expect, but have not proven, that our results can be generalized to multiclass audio segmentation tasks and therefore higher level tasks that seek to explore the ever growing stream of wild audio.

This paper presented an initial discussion with the aim to contribute to a systematic understanding of the challenges of audio classification of consumer-produced videos. For future work, we would like to verify our hypothesis that the "softness" properties of NN contribute to its success on wild audio by comparing to a soft-margin SVM, for example. Furthermore, most of the work that reports results on wild audio is inspired by the speech literature and uses tools originally developed in the speech community. Investigating alternative features instead of MFCCs would therefore be a very valid line of work.

7. REFERENCES

[1] T. Hain and P. C. Woodland, "Segmentation and classification of broadcast news audio," in *Proceedings of ICSLP*, 1998, pp. 2727–2730.

[2] DARPA, "Robust Automatic Transcription of Speech," 2011.

[3] Tim Ng, Bing Zhang, Long Nguyen, Spyros Matsoukas, Xinhui Zhou, Nima Mesgarani, Karel Vesely, and Pavel Matejka, "Developing a Speech Activity Detection System for the DARPA RATS Program," in *Proceedings of Interspeech*, 2012.

[4] Ananya Misra, "Speech/Nonspeech Segmentation in Web Videos," in *Proceedings of Interspeech*, 2012.

[5] G. Schwarz, *The Annals of Statistics*, Institute of Mathematical Statistics, 1978.

[6] Marijn Huijbregts, *Segmentation Diarization and Speech Transcription: Surprise Data Unraveled*, Ph.D. thesis, Universiteit Twente, 2008.

[7] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[8] L. Lu, S. Li, and H.J. Zhang, "Content-based audio segmentation using support vector machines," in *In Proceedings of IEEE ICME 2001*, 2001, pp. 749–752.

[9] Benjamin Elizalde, "Segment and Conquer," 2012, 2nd Multimedia and Vision Meeting in Greater New York Area.

[10] Yoshua Bengio and Yann LeCun, "Scaling Learning Algorithms towards AI," *Large-Scale Kernel Machines*, 2007.

[11] Oriol Vinyals and Suman V. Ravuri, "Comparing Multilayer Perceptron to Deep Belief Network Tandem Features for Robust ASR," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2011.

[12] Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri, "OpenFst: A general and efficient weighted finite-state transducer library," in *Proceedings of the Ninth International Conference on Implementation and Application of Automata, (CIAA 2007)*. 2007, vol. 4783 of *Lecture Notes in Computer Science*, pp. 11–23, Springer, http://www.openfst.org.

[13] NIST, "Rich transcription 2006 spring meeting recognition evaluation plan v2," in *In Rich Transcription 2006 Meeting Recognition Workshop*, 2006.

[14] A. Halevy, P. Norvig, and F. Pereira, "The unreasonable effectiveness of data," *Intelligent Systems, IEEE*, vol. 24, no. 2, pp. 8–12, 2009.