INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Joint Estimation of Articulatory Features and Acoustic models for Low-Resource Languages

Basil Abraham, S. Umesh, Neethu Mariam Joy

Indian Institute of Technology-Madras, India
{ee11d032,umeshs,ee11d009}@ee.iitm.ac.in

Abstract

Using articulatory features for speech recognition improves the performance of low-resource languages. One way to obtain articulatory features is by using an articulatory classifier (pseudo-articulatory features). The performance of the articulatory features depends on the efficacy of this classifier. But, training such a robust classifier for a low-resource language is constrained due to the limited amount of training data. We can overcome this by training the articulatory classifier using a high resource language. This classifier can then be used to generate articulatory features for the low-resource language. However, this technique fails when high and low-resource languages have mismatches in their environmental conditions. In this paper, we address both the aforementioned problems by jointly estimating the articulatory features and low-resource acoustic model. The experiments were performed on two low-resource Indian languages namely, Hindi and Tamil. English was used as the high-resource language. A relative improvement of 23% and 10% were obtained for Hindi and Tamil, respectively.

Index Terms: speech recognition, articulatory features, low-resource languages, deep neural networks (DNN)

1. Introduction

Deep neural network (DNN) based automatic speech recognition systems (ASR) has become the state of the art technique by replacing traditional Gaussian mixture models (GMM) in hidden Markov model (HMM). However, improving DNN performance by combating speaker variations and environmental noise is an active research area. This has generated a lot of interest in developing features that are robust to the aforementioned problems.

In the past, incorporating speech production knowledge using articulatory features was shown to improve ASR performance [1, 2, 3, 4, 5]. Articulatory features (AF) [6] represents speech signal in terms of the underlying attributes of speech production. Kirchoff *et al.* [7] have shown that articulatory features are robust to speaker and channel variations. Articulatory features can be generated in three ways, (i) using physical measurements of the position of the articulators using imaging techniques like cine-radiography [8], (ii) using inverse filtering techniques on acoustic signal [9] and (iii) using articulatory classifiers trained from speech data [7]. The articulators features extracted from articulatory classifiers are often called as pseudo-articulatory features (pseudo-AF) and from now on we refer to them as articulatory features. In this paper, we focus on this method of articulatory feature generation. We use the discrete multi-level feature set [10] as shown in table 1.

Neural network architectures for extracting articulatory features were studied in [7, 11, 10, 12]. In [7], using articulatory features in adverse acoustic conditions were investigated and was shown to significantly improve the recognition performance of ASR. The Johns Hopkins 2006 summer workshop on Articulatory Feature-based Speech Recognition also investigated on the use of articulatory features for ASR [11, 10].

The articulatory features were also used in low-resource scenario in [10, 13, 12]. In this case it is difficult to train a robust articulatory classifiers just with the low-resource data. Hence, the usual practice is to train the articulatory classifiers in another language where data is available (high-resource language) and use these classifiers to extract articulatory features for the low-resource language. This technique was used in [13, 14, 15, 16, 17]. However, this method failed in some cases. For instance, in [17], articulatory features were extracted for Mandarin broadcast news task with articulatory classifiers trained on English continuous telephone speech. The Mandarin acoustic model built with these features failed to perform at par with the model trained on conventional features like filterbank. A similar work in [14] uses articulatory classifiers built with English to extract articulatory features for Hungarian telephone speech. But in this case an improved recognition performance was obtained for the Hungarian acoustic model trained on articulatory features when compared to conventional features. However, when articulatory classifiers were trained from Hungarian data itself, the articulatory features so extracted gave performance improvement over articulatory features extracted from English articulatory classifiers. Both the aforementioned works and [15] attributed the performance degradation to the differences in domain and channel variations between the low-resource and high-resource databases.

In this paper, we address the various issues in using articulatory features for low-resource languages, which are mentioned above. We propose a joint estimation framework where the articulatory features and acoustic model are jointly estimated. In all the earlier works with pseudo-AF, first the articulatory classifiers were trained and these were in-turn used to extract articulatory features for further acoustic modeling. In the proposed approach the articulatory classifiers and the acoustic model are jointly estimated by propagating the cross-entropy error in the final context-dependent states of the acoustic model all the way through the articulatory classifiers. In this paper, we show that joint estimation framework helps in training better articulatory classifiers with just low-resource data. Additionally, we also adapt a well trained articulatory classifier in a high-resource language towards low-resource language. We observed consistent improvements in both cases for two under-resourced Indian languages, Hindi and Tamil.

The paper is organized as follows. A review pseudo articulatory feature extraction is given in section 2. A brief description of the database and experimental setup is given in section 3. The proposed joint estimation framework and the cross-lingual joint estimation is explained in section 4 and 5 respectively. Experimental results are analyzed in detail in section 6. The major contributions of the paper are highlighted in section 7.

http://dx.doi.org/10.21437/Interspeech.2017-1028

Table 1: *Articulatory feature set*

Group	Card-inality	Feature values
Place	10	alveolar (ALV), dental (DEN), labial (LAB), labio-dental (L-D), lateral (LAT), none, post-alveolar (P-A), rhotic (RHO), velar (VEL)
Degree & Manner	6	approximant (APP), closure (CLO), FLAP, fricative (FRIC), vowel (VOW)
Nasality	3	-, +
Rounding	3	-, +
Glottal State	4	aspirated (ASP), voiceless (VL), voiced (VOI)
Vowel shape	23	aa, ae, ah, ao, aw1, aw2, ax, ay1, ay2, eh, er, ey, ey1, ey2, ih, iy, ow1, ow2, oy1, oy2, uh, uw, nil
Height	8	HIGH, LOW, MID, mid-high (MID-H), mid-low (MID-L), very-high (VI), nil
Frontness	7	back (BK), front (FRT), MID, mid-back (MID-B), mid-front (MID-F), nil

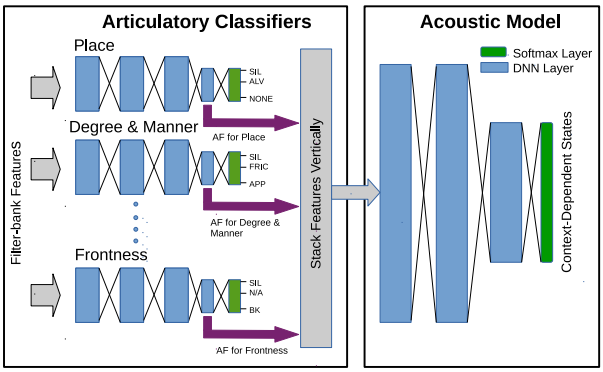

Figure 1: **DNN-sep***: Articulatory classifiers trained separately and are used to extract articulatory features for acoustic modeling.*

2. Review of Articulatory Feature Extraction

In [7, 10, 11], the articulatory features were extracted using the articulatory classifiers as shown in figure 1. To extract articulatory features the articulatory classifiers need to be trained for each of the AF group given in table 1. In the earlier works [7, 10, 11], the articulatory classifiers were trained separately and the features obtained from these classifiers were stacked to generate the final articulatory features. Figure 2 shows the training procedure for building an articulatory classifier for the AF group "Place of Articulation". To train the articulatory classifier for the AF group "Place of Articulation" a neural network classifier is trained with output targets as AF labels in that AF group. Training such a neural network requires the acoustic features to be aligned at frame-level in terms of the corresponding AF labels. Manually transcribing at frame-level in terms of these labels is difficult. Hence the usual practice is to obtain these alignments at frame-level in terms of phones in that language using an acoustic model and map the phones to AF labels

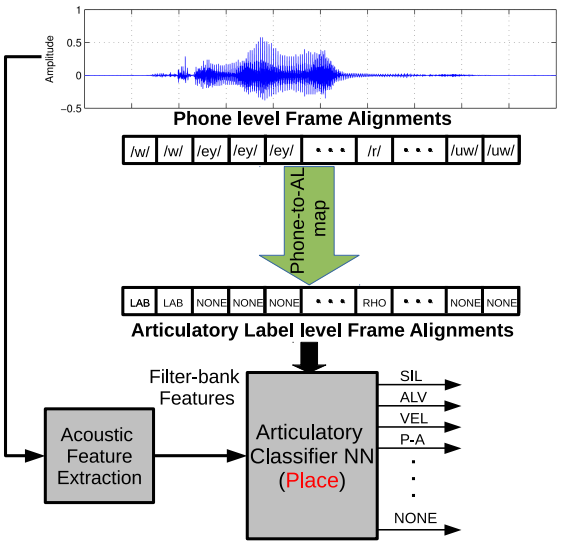

Figure 2: *Block schematic of training an articulatory classifier for place of articulation.*

using a phone-to-AL mapping. Once the articulatory classifiers are trained the features are forward passed through each classifier to obtain the articulatory features. The efficacy of these features depends on the amount of data available to train these classifiers.

3. Database & Experimental Setup

To validate the claim of the proposed techniques the experiments were performed in two under-resourced Indian languages namely, Hindi and Tamil from the MANDI databases. The MANDI database is a multilingual database consisting of 12 Indian languages collected for "Speech-based access to agricultural commodity prices", a Government of India project to build ASR systems in Indian languages to provide price information of agricultural commodities to farmers. The speech was collected form the end users in their work place which vary from quiet to very noisy environments. We used approximately 10 hours of training data and 3-4 hours of test data for both Hindi and Tamil. To perform the cross-lingual experiments 110 hours of Switchboard corpus [18] was used as the English database. To validate the cross-lingual experiments in matched conditions, the SVitchboard task [19] a small vocabulary task defined using the subsets of Switchboard-1 corpus with words ranging from 10 to 500 was used. We refer to this corpus as swbd-6hr.

To build baseline acoustic model we used 13-dimensional MFCC features in GMM-HMM and 40-dimensional filter-bank features for DNN training. The baseline HMM-GMM experiments were performed in Kaldi toolkit [20] to generate the frame-level alignments for DNN training. All the experiments with DNN were performed in CNTK toolkit [21].

4. Proposed Technique of Joint Estimation

In all the previous works using pseudo articulatory features, the articulatory feature extraction and acoustic modeling were performed independently. In this paper a joint estimation framework of estimating the articulatory features and acoustic model is proposed. In the proposed technique the training of the DNN acoustic model propagates the error back even into the articulatory classifiers generating the articulatory features. Hence, the

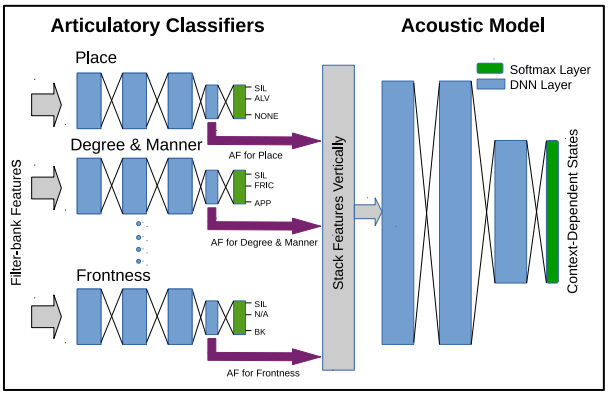

Figure 3: **DNN-joint-est**: *Proposed technique of joint estimation of articulatory classifiers and acoustic model.*

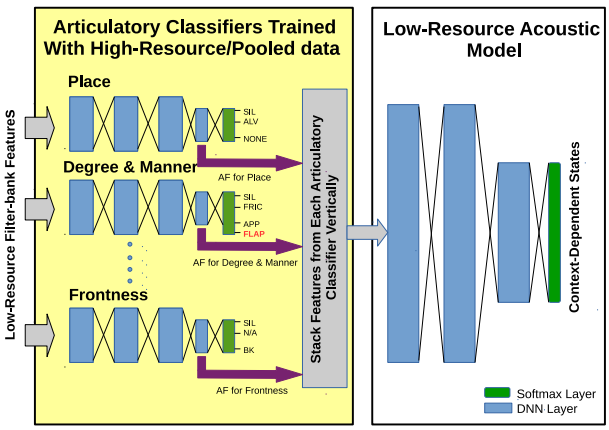

Figure 4: **Cross-DNN-sep**: *Articulatory feature extraction of a low-resource language with articulatory classifiers trained with high-resource/pooled data.*

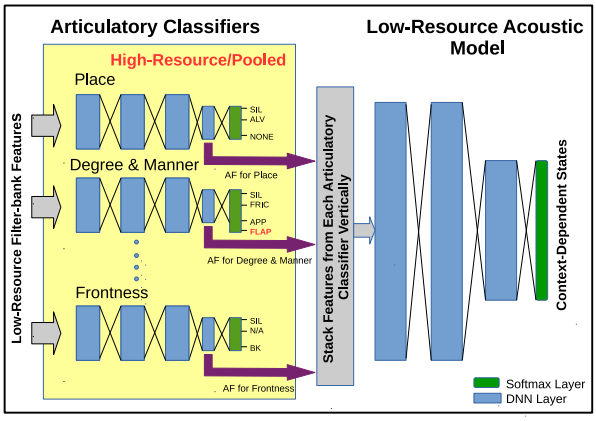

Figure 5: **Cross-DNN-joint-est**: *Joint estimation framework for articulatory classifiers trained with high-resource/pooled data and acoustic model for low-resource language.*

articulatory classifiers also learns to generate the features which makes the phonetic acoustic model perform better.

The block schematic of the proposed joint estimation is given in figure 3. In this, the first stage is the articulatory classifiers which generate the articulatory features for the second stage of acoustic modeling. The entire network is jointly estimated with softmax layers in both stages. The network has a total of 9 softmax layers, one for articulatory classifiers of each AF group and the last softmax layer for the final context-dependent targets of the acoustic model. A softmax layer is required for each articulatory classifier to make sure that each classifier gets trained to generate the corresponding articulatory features during joint estimation. The target for the acoustic model is obtained from context-dependent states of the HMM system and these targets are mapped into the corresponding AF labels in each AF group using the phone-to-AL mapping as shown in 2. The final loss function $L(\theta)$ is a sum of all the individual loss functions as shown below.

$$\begin{aligned} L(\theta) &= \alpha_1 L_{Place}(\theta) + \alpha_2 L_{DM}(\theta) + \alpha_3 L_{Nasality}(\theta) \\ &+ \alpha_4 L_{Rounding}(\theta) + \alpha_5 L_{Glottal}(\theta) + \alpha_6 L_{Vowel}(\theta) \\ &+ \alpha_7 L_{Height}(\theta) + \alpha_8 L_{Frontness}(\theta) \\ &+ \alpha L_{Tied-states}(\theta) \end{aligned}$$

The entire network parameters are learned using back propagation algorithm. The parameters in the acoustic modeling stage learn from the gradient with respect to the cross-entropy error in the final context-dependent state. Whereas, the weight parameters in the articulatory classifiers learn from both the cross-entropy error at the context-dependent states and the error from the articulatory labels belonging to the corresponding AF group. Thus at a point each articulatory classifier weights are updated by considering both these losses and are not affected by the errors occurring at the classifiers belonging to other AF groups.

4.1. Experiments with Joint Estimation

In this case two sets of experiments were performed in both languages. In the first experiment the articulatory classifiers were separately trained on the low-resource language and were used to generate articulatory features for low-resource language as shown in figure 1. We refer this experiment as *DNN-sep* since, DNN's were trained separately. In the second experiment all the eight articulatory classifiers were jointly estimated along with the phonetic acoustic model with context-dependent targets as described in section 4. The block schematic of the proposed joint estimation technique is given in figure 3 and is referred to as *DNN-joint-est*. In the experiment with joint estimation the articulatory classifier weights were initialized from the articulatory classifier weights of the *DNN-sep* model and a random initialization was used for the acoustic modeling stage. Then the entire network was jointly estimated as described in section 4. We have tuned all the α parameters in the loss function and found that using equal weights gave the best recognition performance. We used all the α to be 1. The recognition performance for the DNN-sep and DNN-joint-est experiments are given in table 2.

5. Proposed Technique of Cross-lingual Joint Estimation

A popular approach in extracting articulatory features in low-resource language is to use the well-built articulatory classifiers from a high-resource language as shown in figure 4. However, [13, 14, 15, 16, 17] a degradation in performance was seen in these works due to the mismatch in the environment conditions

Table 2: *Experiments on Articulatory Classifiers Trained only with Low-Resource Language*

	Hindi	Tamil	swbd-6hr
DNN-sep	18.65	18.14	48.21
DNN-joint-est	**16.32**	**17.24**	**46.14**

during data collection. In this paper, we also show that the joint estimation of articulatory classifiers and acoustic model can solve this problem.

The block schematic of the cross-lingual joint estimation is shown in figure 5. In this case also we follow a similar training procedure as described in section 4. However, the main difference is in the estimation of the articulatory classifiers. Here, the articulatory classifiers are trained from high-resource language along with low-resource language. Then this model is used to initialize the articulatory classifiers and then jointly re-estimate with acoustic model using low-resource language data. In the experimental section we will show that pooling gives only a marginal improvement and the major gains are coming from the joint estimation process.

5.1. Experiments with Cross-lingual Joint Estimation

In this section we discuss the case when a high-resource language is available. In this scenario we can train the articulatory classifiers using the high-resource language and use them to generate the articulatory features for the low-resource language. The cross-lingual experiments were performed with *switchboard corpus being high-resource language* and Hindi, Tamil and swbd-6hr as low-resource language. To alleviate the problem of mismatch in data used to train the articulatory classifier and the low-resource language, we use the proposed cross-lingual joint estimation as described in section 5. In this experiment we use a joint estimation setup as shown in figure 5. Here in the first stage of articulatory classifier, the model parameters were initialized from the separately trained articulatory classifiers from each AF group and the acoustic model stage were randomly initialized. Now the entire model was jointly estimated as described in section 5. Here the α weight parameters were kept to unity for both articulatory classifiers and acoustic modeling. In this context we have performed experiments with swbd-6hr (SVitchboard corpus) from the same conditions as of high-resource to confirm our claim that joint estimation helps in mismatched conditions.

The second experiment under cross-lingual scenario was with pooling the data from low-resource language along with high-resource language to build the articulatory classifiers. In this case also we performed experiments with articulatory classifiers being estimated separately and using joint cross-lingual joint estimation. The recognition performances for cross-lingual experiments are given in table 3.

6. Results & Discussion

The recognition performance of the proposed techniques are given in tables 2 and 3. In all the experiments it consistently shows that jointly estimation of articulatory features and acoustic models give improved performance compared to the traditional method of estimating separately.

- **Joint estimation framework with only low-resource data**: In all the three low-resource languages, we obtained consistent improvements with joint estimation of articulatory classifiers and the acoustic model as shown in table 2. This shows that propagating the error from the final context-dependent states of the acoustic modeling stage into the articulatory classifier stage helps in generating better articulatory features.

Table 3: *Experiments on Articulatory Classifiers Trained with High-Resource Language*

Joint-estimation	Pooling	Hindi	Tamil	swbd-6hr
×	×	22.53	22.82	40.81
✓	×	16.57	18.57	**40.09**
×	✓	22.10	22.52	-
✓	✓	**14.21**	**16.39**	-

- **Cross-lingual joint estimation framework**: In this case we can analyze the results under two categories
  - **Matched conditions**: It is the cross-lingual experiments on swbd-6hr with swbd-110hr as high-resource. In this scenario we found that even without joint estimation the acoustic model gave improved performance than DNN-sep experiment with swbd-6hr data alone. The joint estimation framework gave improvements in this case also.
  - **Mismatched condition**: The cross-lingual experiments of Tamil and Hindi with swbd-110hr as high-resource language is considered as the mismatched condition. In this case, the articulatory features extracted from swbd-110hr articulatory classifiers fails to improve the performance of the low-resource acoustic model. However, the cross-lingual joint estimation improves the performance by 4-5% absolute as shown in table 3.
- **Experiments with pooling data**: In the cross-lingual scenario experiments were also performed after pooling data from both high-resource and low-resource language. In the case where articulatory classifiers were separately trained, we did not notice much improved due to mismatched conditions as reported in [12]. However, the joint estimation of this pooled classifiers with low-resource acoustic model improved the performance by 6-7% absolute as shown in table 3.

7. Conclusion

In this work, we proposed a framework to jointly estimate the articulatory classifiers and acoustic model. The proposed method helped in building a robust articulatory classifiers with limited amount of training data. At the same time the performance of the articulatory features were improved with the proposed cross-lingual joint estimation technique under mismatched cross-lingual articulatory classifiers. With the proposed approaches a relative improvement of 23% and 10% is observed in Hindi and Tamil respectively.

8. Acknowledgements

This work was supported in part by the consortium project titled "Speech-based access to commodity price in six Indian languages", funded by the TDIL program of DeITY of Govt. of India. The authors would like to thank consortium members involved in collecting Hindi and Tamil corpus.

9. References

[1] O. Schmidbauer, "Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations," in *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on*. IEEE, 1989, pp. 616–619.

[2] K. Elenius and G. Takács, "Phoneme recognition with an artificial neural network." in *EUROSPEECH*, 1991.

[3] E. Eide, J. R. Rohlicek, H. Gish, and S. Mitter, "A linguistic feature representation of the speech waveform," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2. IEEE, 1993, pp. 483–486.

[4] L. Deng and D. Sun, "Phonetic classification and recognition using hmm representation of overlapping articulatory features for all classes of english sounds," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 1. IEEE, 1994, pp. I–45.

[5] K. Erler and G. H. Freeman, "An hmm-based speech recognizer using overlapping articulatory features," *The Journal of the Acoustical Society of America*, vol. 100, no. 4, pp. 2500–2513, 1996.

[6] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond, and M. Wester, "Speech production knowledge in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 121, no. 2, pp. 723–742, 2007.

[7] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3, pp. 303–319, 2002.

[8] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, J. Zacks, and S. Levy, "Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data," *The Journal of the Acoustical Society of America*, vol. 92, no. 2, pp. 688–700, 1992.

[9] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 1, pp. 133–150, Jan 1994.

[10] J. Frankel, M. Magimai-doss, S. King, K. Livescu, and Ö. Çetin, "Articulatory feature classifiers trained on 2000 hours of telephone speech," in *In Proc. Interspeech*, 2007.

[11] O. Cetin, A. Kantor, S. King, C. Bartels, M. Magimai-Doss, J. Frankel, and K. Livescu, "An articulatory feature-based tandem approach and factored observation modeling," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–645.

[12] Basil Abraham and S. Umesh, "An automated technique to generate phone-to-articulatory label mapping," *Speech Communication*, vol. 86, pp. 107 – 120, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639316300206

[13] S. Sivadas and H. Hermansk, "On use of task independent training data in tandem feature extraction," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, vol. 1. IEEE, 2004, pp. I–541.

[14] L. Tóth, J. Frankel, G. Gosztolya, and S. King, "Cross-lingual portability of mlp-based tandem features–a case study for english and hungarian," 2008.

[15] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multi-stream posterior features for low resource lvcsr systems." in *INTERSPEECH*, 2010, pp. 877–880.

[16] P. Lal and S. King, "Cross-lingual automatic speech recognition using tandem features," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 12, pp. 2506–2515, 2013.

[17] Ö. Çetin, M. Magimai-Doss, K. Livescu, A. Kantor, S. King, C. Bartels, and J. Frankel, "Monolingual and crosslingual comparison of tandem features derived from articulatory and phone mlps," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*. IEEE, 2007, pp. 36–41.

[18] J. Godfrey and E. Holliman, "Switchboard-1 Release 2 LDC97S62," *Linguistic Data Consortium*, 1993.

[19] S. King, C. Bartels, and J. Bilmesy, "SVitchboard 1: small vocabulary tasks from Switchboard," in *Annual Conference of the International Speech Communication Association*, 2005, pp. 3385–3388.

[20] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, December 2011.

[21] D. Yu, A. Eversole, M. Seltzer, K. Yao, O. Kuchaiev, Y. Zhang, F. Seide, Z. Huang, B. Guenter, H. Wang, J. Droppo, G. Zweig, C. Rossbach, J. Gao, A. Stolcke, J. Currey, M. Slaney, G. Chen, A. Agarwal, C. Basoglu, M. Padmilac, A. Kamenev, V. Ivanov, S. Cypher, H. Parthasarathi, B. Mitra, B. Peng, and X. Huang, "An introduction to computational networks and the computational network toolkit," Tech. Rep., October 2014. [Online]. Available: https://www.microsoft.com/en-us/research/publication/an-introduction-to-computational-networks-and-the-computational-network-toolkit/