ARTICULATORY TRAJECTORIES FOR LARGE-VOCABULARY SPEECH RECOGNITION

Vikramjit Mitra[1], Wen Wang[1], Andreas Stolcke[2], Hosung Nam[3], Colleen Richey[1], Jiahong Yuan[4], Mark Liberman[4]

[1]Speech Technology and Research Laboratory, **SRI International**, Menlo Park, CA
[2]**Microsoft Research**, Mountain View, CA
[3]**Haskins Laboratories**, New Haven, CT
[4]**University of Pennsylvania,** Philadelphia, PA
[1]{vmitra, wwang, colleen}@speech.sri.com, [2]anstolck@microsoft.com,
[3]nam@haskins.yale.edu, [4]jiahong@sas.upenn.edu, [4]markyliberman@gmail.com

ABSTRACT

Studies have demonstrated that articulatory information can model speech variability effectively and can potentially help to improve speech recognition performance. Most of the studies involving articulatory information have focused on effectively estimating them from speech, and few studies have actually used such features for speech recognition. Speech recognition studies using articulatory information have been mostly confined to digit or medium vocabulary speech recognition, and efforts to incorporate them into large vocabulary systems have been limited. We present a neural network model to estimate articulatory trajectories from speech signals where the model was trained using synthetic speech signals generated by Haskins Laboratories' task-dynamic model of speech production. The trained model was applied to natural speech, and the estimated articulatory trajectories obtained from the models were used in conjunction with standard cepstral features to train acoustic models for large-vocabulary recognition systems. Two different large-vocabulary English datasets were used in the experiments reported here. Results indicate that employing articulatory information improves speech recognition performance not only under clean conditions but also under noisy background conditions. Perceptually motivated robust features were also explored in this study and the best performance was obtained when systems based on articulatory, standard cepstral and perceptually motivated feature were all combined.

Index Terms— *large vocabulary speech recognition, articulatory trajectories, vocal tract variables, artificial neural networks.*

1. INTRODUCTION

Variability in spontaneous speech adversely affects current state-of-the-art automatic speech recognition (ASR) systems. One major source of variability is coarticulation and reduction; coarticulation is the spreading of features from one segment to another [1]. It has been suggested [2] that variation in speech can be accounted for by incorporating speech production knowledge, which in turn may improve the performance of an ASR system. Several studies [3, 4, 5] have demonstrated that articulatory information can improve the performance of an ASR system by systematically accounting for variability such as coarticulation. Studies [6, 7] have also shown that articulatory information can improve the noise robustness of ASR systems.

In a typical ASR application, the only observable is the speech signal, and speech production knowledge (typically, articulatory information or their dynamic information) is not available. Hence, production information needs to be estimated from the speech signal. Early efforts [8, 9, 10] tried to decipher appropriate features that captured articulatory dynamics and/or events, which were also known as articulatory features (AFs). Schmidbauer [11] proposed an AF-based ASR system, using 19 AFs (describing the manner and place of articulation) to perform phone recognition of German speech and reporting an improvement of 4% over the Mel-frequency cepstral coefficient (MFCC)–Hidden Markov Model (HMM) baseline. Deng *et al.* used 18 multi-valued AFs [12, 13] describing the place of articulation, horizontal - vertical tongue body movement, voicing information and reported a relative phone recognition improvement of 9% over the MFCC-HMM baseline on the TIMIT dataset. A comprehensive literature survey on the use of AFs and production motivated ASR systems was presented in [14].

Articulatory trajectory information is more challenging to use than AFs as it involves retrieving articulatory dynamics from the speech signal, which is traditionally called 'speech inversion'. Inversion studies involving articulatory trajectories have been mostly confined to predicting such dynamics efficiently and accurately, and understanding their functional relationship with the acoustics. Due to the difficulty in estimating them, only a few ASR studies [4, 6, 15] have used such articulatory dynamics. Frankel *et al.*[4] developed a recognition system that uses a combination of acoustic and articulatory information as input, with the articulatory trajectories modeled using phone-specific linear dynamic models. They showed that using articulatory data from direct measurements in addition to MFCCs resulted in a performance improvement of 9% [15] over the system using only MFCCs. Such an improvement did not hold when the articulatory data was estimated from the acoustic signal [15]. A recent study [6] of digit recognition on the Aurora-2 dataset has shown that articulatory trajectories can help to improve the noise robustness of an ASR system.

In this paper we train an artificial neural network (ANN) to estimate articulatory trajectories from a speech signal. We then use it to generate trajectories that are used as features for training and testing English large-vocabulary continuous speech recognition (LVCSR) systems. Our results indicate that use of articulatory information in addition to standard cepstral features provides sufficient complementary information that helps to reduce the word error rates (WERs) in both clean and noisy conditions.

2. DATASETS

To train a model for estimating vocal tract constriction variable trajectories (a.k.a TVs) from speech, we require a speech database containing ground-truth TVs. Unfortunately; no such database is exist at present. For this reason, Haskins Laboratories' Task

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

Dynamic model (popularly known as TADA [16]) along with HLsyn [17] was used in our work to generate a database that contains synthetic speech along with articulatory specifications. From text input, TADA generates vocal tract constriction variables and other parameters, some of which are used by HLsyn to create the corresponding synthetic speech. TVs (refer to [6, 7] for more details) are continuous time functions that specify the shape of the vocal tract in terms of constriction degree and location of the constrictors. TADA defines eight TVs altogether as shown in Table 1, and their positional information is pictorially represented in Fig. 1. Refer to [18-20] for more explanation about the TVs..

Table 1. Constriction organ, vocal tract variables, their unit of measurement and dynamic range

Constriction organ	Vocal tract variables	*Unit*	*Dynamic range*	
			Max	*Min*
Lip	Lip Aperture (LA)	mm	27.00	-4.00
	Lip Protrusion (LP)	mm	12.00	8.08
Tongue Tip	Tongue tip constriction degree (TTCD)	mm	31.07	-4.00
	Tongue tip constriction location (TTCL)	degree	80.00	0.00
Tongue Body	Tongue body constriction degree (TBCD)	mm	12.50	-2.00
	Tongue body constriction location (TBCL)	degree	180.00	87.00
Velum	Velum (VEL)	-	0.20	-0.20
Glottis	Glottis (GLO)	-	0.74	0.00

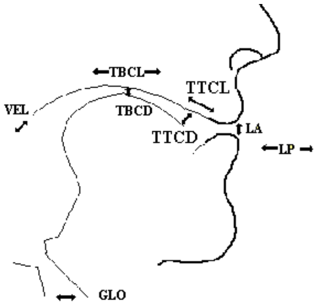

Fig. 1. Eight tract variables from five distinct constriction locations

From the CMU dictionary [30] 111,929 words were selected and their Arpabet pronunciations were input to TADA, which generated their corresponding TVs (refer to Table 1) and synthetic speech. 80% of the data was used as the training set, 10% was used as the development set, and the remaining 10% was used as the test set. Note that TADA generated speech signals at a sampling rate of 8 kHz and TVs at a sampling rate of 200 Hz.

For LVCSR experiments we used two datasets – (1) Aurora-4 [21] and (2) 400 hours Fisher subset (fsh2004sub) [22] with NIST RT-04 Conversational Telephone Speech development set as the test set (denoted dev2004).

Aurora-4 contains six additive noise versions with channel matched and mismatched conditions. It is created from the standard 5K Wall Street Journal (WSJ0) database and has 7180 training utterances of approximately 15 hours duration and 330 test utterances. The acoustic data (both training and test sets) comes with two different sampling rates (8 kHz and 16 kHz). In Aurora-4, two training conditions were specified: (1) clean training, which is the full SI-84 WSJ training set without added noise; and (2) multi-condition training, with about half of the training data recorded using one microphone, and the other half recorded using a different microphone, with different types of added noise at different signal-to-noise ratios (SNRs). The noise types are similar to the noisy conditions in the test set. The Aurora-4 test data includes 14 test sets from two different channel conditions and six different added noises (in addition to the clean condition). The SNR was randomly selected between 0 and 15 dB for different utterances. The six noise types used were (1) car, (2) babble, (3) restaurant, (4) street, (5) airport and (6) train-station. The evaluation set comprised 5K words under two different channel conditions. The original audio data for test conditions 1-7 was recorded with a Sennheiser microphone, while test conditions 8-14 were recorded using a second microphone that was randomly selected from a set of 18 different microphones (more details in [21]). The different noise types were digitally added to the clean audio data to simulate noisy conditions.

The 400 hour Fisher corpus (a.k.a. fsh2004sub) [22] is a subset of the LDC Fisher data and contains conversational English speech and a balance of speaker gender, conversational topics, and phone line conditions. For acoustic models trained with fsh2004sub, the test set was the NIST RT-04 Conversational Telephone Speech development set (a.k.a. dev2004). No noise is intentionally added to any of the files in these datasets.

3. TV ESTIMATOR

ANNs have been used [6, 23] for estimating TV trajectories from the speech signal. Once trained, ANNs require low computational resources compared to other methods in terms of both memory requirements and execution speed [23]. ANN has the advantage that it can have *M* inputs and *N* outputs; hence, a complex mapping of *M* vectors into *N* different functions can be achieved. In such architecture, the same hidden layers are shared by all *N* outputs, endowing the ANN with the implicit capability to exploit any correlation that the N outputs may have amongst themselves. The feed-forward ANNs were trained with back propagation using a scaled conjugate gradient (SCG) algorithm. For TV estimation, the speech signal was parameterized as Normalized Modulation Cepstral Coefficients (NMCCs) [24], where 13 cepstral coefficients were extracted at the rate of 100 Hz with an analysis window of 20 ms. The TVs were downsampled to 100 Hz to temporally synchronize with the NMCCs. The NMCCs and TVs were Z-normalized and scaled to fit their dynamic ranges into [-0.97, +0.97]. It was observed [23] that incorporating dynamic information helps to improve the speech-inversion performance, for which the input features were contextualized by concatenating every other feature frame within a 200 ms window. Dimensionality reduction was performed on each feature dimension by using discrete cosine transform (DCT) and retaining 70% of the coefficients, resulting in a feature of dimension 104. Hence, for the TV estimator *M* was 104 and *N* was 8 for the eight TV trajectories.

4. ACOUSTIC MODEL

The details of all algorithms used for building the recognition systems in this work were described in [25]. The Aurora-4 system uses a bigram language model (LM) on the initial pass and uses second-pass decoding with model space maximum likelihood linear regression (MLLR) speaker adaptation followed by trigram LM rescoring of the lattices from the second pass. We trained four sub-systems: (1) MFCC system: using SRI's Decipher front end MFCC features, (2) RASTA-PLP system: using perceptual linear prediction with RASTA-filtering, (3) NMCC system: using perceptually motivated NMCC acoustic features [24], and (4)

MFCC+ModTV_pca30 system: using MFCC features combined with the temporal modulation information of the estimated TVs, which were then reduced to 30 dimensions (D) using principal component analysis (PCA). Our initial experiments revealed that using temporally contextualized TVs as features provided better ASR performance than using the instantaneous TVs, indicating that the dynamic information of the TVs contributes to improving ASR performance. A context of 13 frames i.e., ~120 ms of temporal information was used to contextualize the TVs. To reduce the dimension of the contextualized TVs, DCT was performed on each of the eight TV dimensions and their first seven coefficients were retained, resulting in a 56D feature. We name this feature the modulation of TVs (ModTVs), which were concatenated with 52D MFCCs (in all experiments the standard cepstral features contained 13 cepstral coefficients along with velocity, acceleration, and jerk information). The resulting 108D feature set was PCA transformed to a 30D feature set, as we observed that more than 90% of the information resided in the Eigenvector space of the first 30 Eigenvalues.

For the fsh2004sub system, the LMs were pre-trained and include bigrams (for lattice generation) and higher-order LMs (for lattice and N-best rescoring). For training the LMs, the CTS in-domain transcripts are augmented with data harvested from the web, using a search engine to select data that is matched for both style and content. As with the Aurora-4 experiments, we built multiple subsystems using different front-end features: (1) MFCC, (2) RASTA-PLP, (3) NMCC and (4) MFCC+ModTV_pca30, where the baseline system used 52D MFCC features.

The features were normalized using standard cepstral mean and variance normalization on the Fisher speaker clusters (conversation sides). For MFCC and RASTA-PLP features, we conducted vocal tract length normalization (VTLN). For MFCC, RASTA-PLP and NMCC heteroscedastic linear discriminant analysis (HLDA) was used to reduce the 52D features to 39D. HLDA was not used for MFCC+ModTV_pca30. For all front-end features, we trained maximum likelihood estimate (MLE) cross-word HMM-based acoustic models with decision-tree clustered states.

When decoding the testsets, the cross-word MLE model was first adapted through MLLR using a phone-loop model as reference and then used for 1-best decoding. The cross-word MLE model was adapted again through MLLR on the 1-best decoding output and the adapted model and the bigram LM were used for generating HTK lattices. Speaker-clustered regression class trees were used to improve robustness of MLLR adaptation [26]. We then dumped 2000-best N-best lists from the HTK lattices and rescored with the 4-gram LM. The final output is the result of an M-way combination of the subsystems based on different front end features, using N-best ROVER [27] implemented in SRILM [28].

5. RESULTS

For the TV estimator, a single hidden layered feed-forward ANN with tan-sigmoid activation was used and it was observed that a network with 150 neurons with biases provided reliable results. Performance of the TV estimator was measured using Pearson's product-moment correlation (PPMC) coefficient (r_{PPMC}) between the actual or ground-truth and the estimated articulatory trajectories. Table 2 gives the r_{PPMC} on the test set of the TADA synthesized corpus. Note that the ANN based TV-estimator was trained and tested on the synthetic dataset created using TADA. Previous studies [23] on TV estimation have reported average r_{PPMC} values of ~0.94 using a smaller corpus of lesser than 500 words. The corpus used in this paper is more than 200 times larger than what was used in [23], and r_{PPMC}~0.93 suggests that the TV estimator is performing similar to the systems presented elsewhere. The TV estimator was then deployed on the natural speech of Aurora-4, fsh2004sub and dev2004, and the estimated TVs obtained were used in the ASR experiments reported below.

Table 2. *r_{PPMC} for each TV obtained from the ANN*

GLO	VEL	LA	LP	TTCD	TTCL	TBCD	TBCL
0.938	0.948	0.897	0.908	0.919	0.921	0.920	0.955

For the Aurora-4 LVCSR experiments we used only mismatched conditions (i.e., train with clean data [clean training] and test on data from different noisy background and the same or different channels) at 8 kHz. We explored four different feature sets: (1) MFCCs; (2) RASTA-PLP from SRI International's DECIPHER® front end; (3) NMCC; and (4) MFCC+ModTV_pca30. Tables 3 and 4 show WERs for the 8 kHz clean training condition. Note that we have also explored fusing ModTVs with RASTA-PLP and NMCCs, but such fusion didn't to provide any gain. Table 3 represents the result from matched channel conditions, with the training and test files representing identical channel conditions. Table 4 represents mismatched channel conditions, in which test files represent a different channel condition than the training files. Results in both Table 3 and 4 are from using a bigram LM in decoding. Tables 5 and 6 present N-best ROVER results after rescoring N-best lists from each of the four different subsystems with a trigram LM. Tables 7 and 8 present the M-way ROVER combination of trigram rescored n-best lists for the above four systems. As observed from the tables below, trigram rescoring and N-best ROVER helped to reduce the WERs significantly compared to the original bigram decoding. An M-way ROVER combination further improved the performance, by utilizing complementary information amongst the different systems. ROVER [29] combines 1-best outputs from multiple ASR systems to produce a composite output having lower WER. In the matched-channel condition, M-way n-best ROVER combination reduces the average WER from 35.2% in the baseline MFCC system to 26.1%.

Table 3. Bigram decoding WER (%) for clean training conditions (with testing channel the same as the training data) at 8 kHz

		MFCC	RASTA-PLP	NMCC	MFCC+ModTV-pca30
1	Clean	14.6	14.2	16.1	14.8
2	Car	20.0	22.2	21.0	22.8
3	Babble	43.6	47.1	37.5	40.3
4	Restaurant	46.4	46.0	41.7	42.6
5	Street	51.0	52.7	40.2	42.1
6	Airport	38.4	38.8	36.8	36.1
7	Train station	50.7	53.1	41.6	46.8
	Average(2-7)	41.7	43.3	36.5	38.3

Table 4. Bigram decoding WER (%) for clean training conditions (with testing channel different from the training data) at 8 kHz

		MFCC	RASTA-PLP	NMCC	MFCC+ModTV-pca30
1	Clean	17.9	20.5	19.5	18.6
2	Car	25.0	30.1	25.3	27.3
3	Babble	49.5	54.5	42.4	49.7
4	Restaurant	53.3	56.9	48.6	53.2
5	Street	57.5	63.0	48.0	57.3
6	Airport	43.3	47.7	42.9	45.5
7	Train station	54.9	59.0	46.7	56.9
	Average(2-7)	47.3	51.9	42.3	48.3

Under mismatched-channel condition, the reduction on WER is from the baseline 41.3% to 31.9% (here adding RASTA-PLP to 3-way combination slightly hurts the performance). For the acoustic models trained with fsh2004sub data and tested with dev2004 data, the original system using bigram LM for decoding and the N-best ROVER performance after rescoring N-best list with the 4-gram LM is shown in Table 9. Note that in Table 9, the MFCC and RASTA-PLP systems both had VTLN performed on the acoustic features, whereas no VTLN was performed on the MFCCs of the MFCC+ModTV-pca30 systems as similar transformation cannot be applied on the ModTVs. We observed that with VTLN an absolute 3% reduction in WER is seen for the MFCCs relative to not using VTLN [MFCC(woVTLN)]. Adding the ModTVs to MFCCs reduced the WER by 2% absolute (from 32.5 to 30.5) in the N-best ROVER systems. M-way ROVER combination was performed on the 4-gram LM rescored n-best lists for the system shown above, and the results are given in Table 10.

Table 5. N-best ROVER WER (%) (after trigram LM rescoring) for clean training conditions (with identical training and testing channels)

		MFCC	RASTA-PLP	NMCC	MFCC+ModTV-pca30
1	Clean	9.8	9.4	11.3	9.8
2	Car	13.8	15.7	14.3	14.6
3	Babble	36.7	39.9	30.9	32.8
4	Restaurant	40.6	39.3	35.3	36.2
5	Street	44.5	46.6	34.1	35.3
6	Airport	30.8	31.4	30.0	29.7
7	Train station	44.5	47.5	34.7	39.7
	Average(2-7)	35.2	36.7	29.9	31.4

Table 6. N-best ROVER WER (%) (after trigram LM rescoring) for clean training conditions (with different and testing channels)

		MFCC	RASTA-PLP	NMCC	MFCC+ModTV-pca30
1	Clean	12.3	14.6	13.8	12.7
2	Car	17.7	22.9	18.0	18.4
3	Babble	43.0	48.8	35.9	37.9
4	Restaurant	47.5	51.5	42.3	41.9
5	Street	53.1	58.0	42.0	45.5
6	Airport	36.4	41.0	36.7	34.8
7	Train station	49.9	53.6	40.3	44.1
	Average(2-7)	41.3	46.0	35.9	37.1

Table 7. M-way ROVER combination (after trigram LM rescoring) for clean training conditions (with identical training and testing channels)

		[MFCC] - [RASTA-PLP]	[MFCC]-[MFCC+ModTV-pca30]	[MFCC]-[MFCC+ModTV-pca30]-[NMCC]	[MFCC]-[MFCC+ModTV-pca30]-[NMCC]-[RASTA-PLP]
1	Clean	8.4	8.6	8.1	7.8
2	Car	12.8	12.5	11.3	11.1
3	Babble	34.8	31.5	27.5	27.6
4	Restaurant	36.0	34.0	29.7	29.9
5	Street	42.1	35.5	30.6	30.7
6	Airport	27.6	26.9	24.5	23.8
7	Train station	42.4	38.3	33.4	33.7
	Average(2-7)	32.6	29.8	26.2	**26.1**

Table 8. M-way ROVER combination (after trigram LM rescoring) for clean training conditions (with different training and testing channels)

		[MFCC] - [RASTA-PLP]	[MFCC]-[MFCC+ModTV-pca30]	[MFCC]-[MFCC+ModTV-pca30]-[NMCC]	[MFCC]-[MFCC+ModTV-pca30]-[NMCC]-[RASTA-PLP]
1	Clean	11.7	10.9	10.6	10.1
2	Car	17.6	16.3	15.0	15.0
3	Babble	42.5	36.8	32.2	33.5
4	Restaurant	46.2	40.8	37.2	37.3
5	Street	51.6	45.3	39.0	40.0
6	Airport	35.7	32.4	29.5	29.2
7	Train station	48.8	43.2	38.2	38.6
	Average(2-7)	40.4	35.8	**31.9**	32.3

Table 10 shows that the M-way ROVER combination between [MFCC]-[MFCC+ModTV-pca30] systems demonstrated lower WER compared to [MFCC]-[RASTAPLP] systems indicating that articulatory features are bringing in more complementary information than the RASTAPLP system. The combination of the perceptually motivated NMCC features further reduced the WER by 1% absolute and finally the best number was obtained by combining all the systems together, with a final WER of 26.6%, resulting in an absolute 2.6% reduction in WER compared to the N-best ROVER result of 29.2% from the MFCC-baseline.

Table 9. Original system and the N-best ROVER (after 4-gram LM rescoring) WER for systems trained with fsh2004sub and tested on dev2004

		Original system	N-best ROVER
Feature-based systems	MFCC	29.9	29.2
	RASTA-PLP	32.0	31.3
	NMCC	33.7	33.0
	MFCC+ModTV-pca30	33.5	30.5
	MFCC(woVTLN)	33.2	32.5

Table 10. WER from M-way ROVER (after 4-gram LM rescoring) combination of different systems trained with fsh2004sub and tested on dev2004

	M-way ROVER
[MFCC]-[RASTA-PLP]	28.7
[MFCC]-[MFCC+ModTV-pca30]	28.3
[MFCC]- [MFCC+ModTV-pca30]-[NMCC]	27.2
[MFCC]-[MFCC+ModTV-pca30]-[NMCC]-[RASTA-PLP]	**26.6**

6. CONCLUSION

We have presented LVCSR experiments using articulatory features, and compared the performance with a standard MFCC baseline, a RASTAPLP system, and perceptually motivated NMCC features. It was observed that the articulatory features always helped to improve performance over the MFCC baseline and the M-way N-best ROVER combination of all the systems demonstrated the best WER. The uniqueness of the study described in this paper lies in the fact we presented three broad class LVCSR front ends (standard cepstral based, speech perception based, and articulatory information based systems) and showed that they capture complementary information and can significantly improve ASR performance when fused together. Performance of the TV estimator discussed in this paper can be further improved by incorporating sophisticated modeling techniques (such as deep learning), which in turn can improve performance of the articulatory feature based LVCSR system, a future direction for research in this area.

7. ACKNOWLEDGMENT

This research was supported by NSF Grant # IIS-0964556, IIS-1162046 and IIS-1161962.

8. REFERENCES

[1] R. Daniloff and R. Hammarberg, “On defining coarticulation”, *J. of Phonetics*, Vol.1, pp. 239-248, 1973.

[2] K. N. Stevens, “Toward a model for speech recognition”, *J. of Acoust. Soc. Am.*, Vol.32, pp. 47-55, 1960.

[3] K. Kirchhoff, *Robust Speech Recognition Using Articulatory Information*, PhD Thesis, University of Bielefeld, 1999.

[4] J. Frankel and S. King, “ASR - Articulatory Speech Recognition”, *Proc. of Eurospeech*, pp. 599-602, Denmark, 2001.

[5] L. Deng and D. Sun, “A statistical approach to automatic speech recognition using atomic units constructed from overlapping articulatory features”, *J. of Acoust. Soc. Am.*, 95(5), pp. 2702-2719, 1994.

[6] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman and L. Goldstein, “Articulatory information for noise robust speech recognition”, *IEEE Trans. on Audio, Speech and Language Processing*, Vol. 19, Iss. 7, pp. 1913-1924, 2010.

[7] V. Mitra, H. Nam and C. Espy-Wilson, “Robust speech recognition using articulatory gestures in a Dynamic Bayesian Network framework”, *Proc. of Automatic Speech Recognition & Understanding Workshop*, ASRU, pp. 131-136, Hawaii, 2011.

[8] R. Cole, R.M. Stern and M.J.Lasry, “Performing Fine Phonetic Distinctions: Templates versus Features”, in *Invariance and Variability of Speech Processes*, edited by J.S. Perkell and D. Klatt (Lawrence Erlbaum Assoc., Hillsdale, NJ), Chap.15, pp. 325-345, 1986.

[9] B. Lochschmidt, “Acoustic-phonetic analysis based on an articulatory model”, *Automatic Speech Analysis and Recognition*, J.P. Hayton (D. Reidel, Dordrecht) eds., pp. 139-152, 1982.

[10] R. D. Mori, P. Laface and E. Piccolo, “Automatic detection and description of syllabic features in continuous speech”, *IEEE Trans. on Acoust., Speech & Sig. Processing*, 24(5), pp. 365-379, 1976.

[11] O. Schmidbauer, “Robust statistic modelling of systematic variabilities in continuous speech incorporating acoustic-articulatory relations”, *Proc. of ICASSP*, pp. 616-619, 1989.

[12] L. Deng and D. Sun, “A statistical approach to ASR using atomic units constructed from overlapping articulatory features”, *J. of Acoust. Soc. Am.*, 95, pp. 2702–2719, 1994.

[13] K. Erler and L. Deng, “Hidden Markov model representation of quantized articulatory features for speech recognition”, *Comp., Speech & Lang.*, Vol. 7, pp. 265–282, 1993.

[14] S. King, J. Frankel, K. Livescu, E. McDermott, K. Richmond and M. Wester, “Speech production knowledge in automatic speech recognition”, *J. of Acoust. Soc. Am.*, 121(2), pp. 723-742, 2007.

[15] J. Frankel, K. Richmond, S. King and P. Taylor, “An automatic speech recognition system using neural networks and linear dynamic models to recover and model articulatory traces”, *Proc. of ICSLP*, Vol. 4, pp. 254-257, 2000.

[16] H. Nam, L. Goldstein, E. Saltzman and D. Byrd, “Tada: An enhanced, portable task dynamics model in Matlab”, *J. of Acoust. Soc. Am.*, 115(5), pp. 2430, 2004.

[17] H. M. Hanson and K. N. Stevens, “A quasiarticulatory approach to controlling acoustic source parameters in a Klatt-type formant synthesizer using HLsyn”, *J. of Acoust. Soc. Am.*, 112(3), pp. 1158-1182, 2002.

[18] C. P. Browman and L. Goldstein, “Towards an Articulatory Phonology”, *Phonology Yearbook*, 85, pp. 219-252, 1986.

[19] C. P. Browman and L. Goldstein, “Articulatory Phonology: An Overview”, *Phonetica*, 49, pp. 155-180, 1992.

[20] E. Saltzman and K. Munhall, “A Dynamical Approach to Gestural Patterning in Speech Production”, *Ecological Psychology*, 1(4), pp. 332-382, 1989.

[21] G. Hirsch, “Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task”, *ETSI STQ-Aurora DSR Working Group*, June 4, 2001.

[22] G. Evermann, H.Y. Chan, M.J.F. Gales, B. Jia, D. Mrva, P.C. Woodland and K.Yu, “Training LVCSR systems on thousands of hours of data”, *Proc. of ICASSP*, vol. 1, 209-212, 2005.

[23] V. Mitra, H. Nam, C. Espy-Wilson, E. Saltzman and L. Goldstein, “Retrieving tract variables from acoustics: a comparison of different machine learning strategies”, *IEEE Journal of Selected Topics on Signal Processing*, Sp. Iss. on Statistical Learning Methods for Speech and Language Processing, Vol. 4, Iss. 6, pp. 1027-1045, 2010.

[24] V. Mitra, H. Franco, M. Graciarena and A. Mandal, “Normalized amplitude modulation features for large vocabulary noise-robust speech recognition”, *Proc. of ICASSP*, pp. 4117-4120, Japan, 2012.

[25] A. Stolcke, B. Chen, H. Franco, V. R. R. Gadde, M. Graciarena, M.-Y. Hwang, K. Kirchhoff, A. Mandal, N. Morgan, X. Lin, T. Ng, M. Ostendorf, K. Sonmez, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng and Q. Zhu, “Recent Innovations in Speech-to-Text Transcription at SRI-ICSI-UW”, *IEEE Trans. on Audio, Speech and Language Processing*, 14(5), pp. 1729-1744, 2006.

[26] A. Mandal, M. Ostendorf and Andreas Stolcke, “Improving robustness of MLLR adaptation with speaker-clustered regression class trees”, *Computer Speech & Language*, 23, pp. 176 199 (2009). ISSN 0885-2308.

[27] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, J. Zheng, “The SRI March 2000 Hub-5 Conversational Speech Transcription System”, Proc. NIST Speech Transcription Workshop, College Park, MD, 2000.

[28] A. Stolcke, “SRILM - An Extensible Language Modeling Toolkit”, *Proc. of ICSLP*, pp. 901-904, 2002.

http://www.speech.sri.com/projects/srilm/.

[29] J. G. Fiscus, “A post-processing system to yield reduced word error rates: Recognizer output voting error reduction. (ROVER),” Proc. of ASRU 1997, pp. 347-354, 1997.

[30] http://www.speech.cs.cmu.edu/cgi-bin/cmudict