2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics
International Conference (PRASA-RobMech)
Port Elizabeth, South Africa, November 26-27, 2015

# Synthetic triphones from trajectory-based feature distributions

Jaco Badenhorst[1,2] and Marelie H. Davel[1,3]
[1]Human Language Technology Research Group, CSIR Meraka, South Africa.
[2]Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa.
[3]CAIR, CSIR Meraka, South Africa.
jacbadenhorst@gmail.com, marelie.davel@gmail.com

***Abstract*— We experiment with a new method to create synthetic models of rare and unseen triphones in order to supplement limited automatic speech recognition (ASR) training data. A trajectory model is used to characterise seen transitions at the spectral level, and these models are then used to create features for unseen or rare triphones. We find that a fairly restricted model (piece-wise linear with three line segments per channel of a diphone transition) is able to represent training data quite accurately. We report on initial results when creating additional triphones for a single-speaker data set, finding small but significant gains, especially when adding additional samples of rare (rather than unseen) triphones.**

**Index Terms**: synthetic triphones, trajectory modelling, trajectory-based features, feature distributions, feature construction

## I. Introduction

The accurate modelling of co-articulation effects in automatic speech recognition (ASR) systems has been a driving force behind the development of large speech corpora [1]. Whole word (or even phrasal) units capture co-articulation effects accurately within unit but require very large training corpora; limited training data forces the use of smaller units, and has resulted in the widespread use of context-dependent phones to capture co-articulation effects [1].

In practice, context sizes of three (triphones) or five (quinphones) are often used. When data is limited, many of these context-dependent units will rarely or ever be seen during training. In typical ASR systems, such unseen context-dependent units are modelled by clustering them with 'matching' seen units, based on a combination of acoustic and linguistic analysis, which is not always an optimal solution [2]. We are interested in determining whether it is possible to generate synthetic versions of such unseen or rare contexts from less specialised units observed in the training data.

First, we require a model that links more general units to more specialised units. For this purpose, we use a trajectory model that provides a compact way of representing the characteristic behaviour of transitions. From the characteristic trajectory behaviour of the less specialised transitions we reconstruct models for unseen transitions. In the current study, we restrict ourselves to triphone modelling, and aim to generate synthetic triphones from seen diphones. If this is possible, the same approach should be applicable to larger contexts, and possibly also to synthesizing additional speech data based on a small sample of data from a given speaker.

## II. Background

In recent work, there has been renewed interest in data augmentation approaches for improving recognition accuracy for under-resourced languages. A useful way to group data augmentation schemes is to consider what type of additional data a technique produces. In [3], the three data types are referred to as 'other language', 'unsupervised' and 'synthesised' data. To incorporate other language data, systems utilise multilingual acoustic models trained on universal phone sets [4] exploiting resources across language barriers. Bootstrapping and filtering out the poor quality data is helpful (unsupervised techniques), while synthetic data may refer to perturbed data or entirely new examples of contexts which have been artificially generated. These techniques can potentially also generate vast amounts of data.

The work of both Jaitly and Hinton [5] and Kanda *et al.* [6] have shown that modelling accuracy can be improved by augmenting limited training data with synthetic training samples. For both cases, a modified version of the training data is added to the original data set when training hidden Markov models with deep neural networks. In the first case, vocal tract length normalisation (VTLN) is applied with different warping factors (the features are adjusted, labels kept unchanged) and in the second, different VTLN warping factors, different speech rates and frequency distortions are applied in a similar fashion.

Using trajectory models for the same goal, builds on prior work analysing co-articulation trajectories [7], [8], [9] as well as various studies on trajectory modelling for ASR purposes [10], [11], [12], [13]. Particularly, in [8] it was found that some decision-tree clustered triphones provided less accurate representations than a simple biphone model, providing the motivation for the current study.

As trajectory models, in effect, smooth features at frame level, they are related to the low-pass filtering used in noise robust speech recognition. The end goal of noise robust approaches is to systematically 'recover' corrupted speech frames. In principle, if the reliable features can be identified, these can then in turn be used to make more accurate predictions about less reliable ones. To this end, Chen and Bilmes [14] use Auto-Regressive Moving Average (ARMA) filtering at the cepstral level to improve ASR robustness in noisy conditions. If over-smoothing occurs, the more definite boundaries of speech events can be modelled using edge-

978-1-4673-7450-7/15/$31.00 ©2015 IEEE

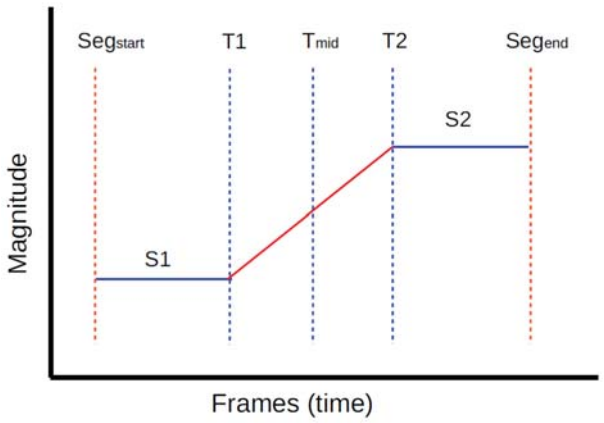


Fig. 1. *Characteristic representation of a single transition (3-piece linear model).*

preserved filtering [15]. Xiao and Li [16] show that besides normalising the probability distributions of speech features, the temporal characteristics of the feature trajectories can also be enhanced at the spectral level. In this work, we apply ARMA filtering at the spectral level, as a preprocessing step prior to fitting trajectories.

## III. Approach

Generating synthetic triphones consists of four main steps: (1) fitting a trajectory model to seen transitions, (2) estimating trajectory parameters based on these seen transitions, (3) creating artificial utterances based on the estimated trajectory parameters, and (4) constructing trajectory-based features for model training.

### A. Diphone segment-based trajectory model

We model the transitions that occur in speech with piecewise linear approximation at the spectral level. Three line pieces are used to fit a single feature channel of a filterbank, using least-squares optimisation. Figure 1 depicts this modelling strategy. A segment effectively describes a diphone, only using the closest $50\%$ of monophone frames to the ASR boundary. We restrict the start and end line segments to be constant values (linear with zero slope), and model the transition between these two values with a straight line of variable slope. We require the constant line segments (the start and end line pieces, referred to as *stable values*) to be associated with at least 1 frame each. The connecting central line segment is referred to as the *change descriptor*.

Each stable value is estimated as the mean of associated feature values; the change descriptor is modelled by the first order line connecting the stable value anchor points. We optimise the squared error ($SE$) across all three line segments simultaneously by searching through the indexes of the possible start and end points for the change descriptor and draw the first order line between the end and starting indexes of the two anchor points. The squared errors at each instant are estimated, followed by the channel-specific mean square error across frames:

$$MSE_{channel}(c) \quad = \quad \frac{1}{F}\sum_{f=1}^{F}|t_c(f) - x_{c,f}|^2 \qquad (1)$$

where $t_c(f)$ is the value of the trajectory function and $x_{c,f}$ the true feature value, respectively, at frame $f$ and feature channel $c$, and $|t_c(f) - x_{c,f}|^2$ is the squared residual. $F$ denotes the total number of frames for the segment. Once optimised, this model provides the following scalar values (see Figure 1):

| | |
|---|---|
| S1, S2 | parameter value at initial and final stable value |
| $T1, T2$ | frame at start and end of the transition |
| $T_{mid}$ | centre of the transition |
| $T_{dur}$ | difference between T2 and T1 |

(2)

A similar SE measurement is also used to evaluate the extent to which trajectory models fit a set of speech data: the mean error (over all frames of all transitions) is now taken across all channels as well. Since channels have quite different standard deviations ($\sigma_c$), a variance-weighted MSE ($MSE_{weighted}$) is useful to evaluate:

$$MSE_{weighted} \quad = \quad \frac{1}{C}\sum_{c=1}^{C}\frac{1}{\sigma_c^2}MSE_{channel}(c) \qquad (3)$$

with $C$ the total number of feature channels.

### B. Predicting trajectory-based parameters

Given a set of training data, any group of segments can be modelled by a probability density function (pdf) over the parameters of section III-A. In this work, we choose to make the assumption that each parameter is normally distributed. The following section now describes how we use pdfs of the segment-based trajectory parameters to model speech data.

*1) Estimating parameter distributions:* Once an initial set of trajectories has been fitted to the training data, the mean and full-covariance matrix is estimated for the stable values ($S1$ and $S2$ respectively) of every particular biphone context that is required. Although biphone contexts are better resourced than triphone contexts, it is still not guaranteed that all biphones will have been seen. To supplement the under-estimated variances of these biphone values, we share the monophone diagonal variances for each distribution estimated on less than a fixed number of examples (3 in the current work).

Time alignments are modelled in a similar manner. Referring back to the schematic representation in Figure 1, these parameters are captured in two pdfs: (1) $T_{dur}$ and (2) $T_{mid}$. A diphone context size is used.

*2) Creating synthetic phones:* As described above, we represent speech data at the spectral level with a set of four full-covariate pdfs (representing $S1$, $S2$, $T_{mid}$ and $T_{dur}$) per modelled context (currently diphones) and filterbank channel. For every triphone synthesised, the two diphones are created individually, by sampling from these four pdfs. Analytic

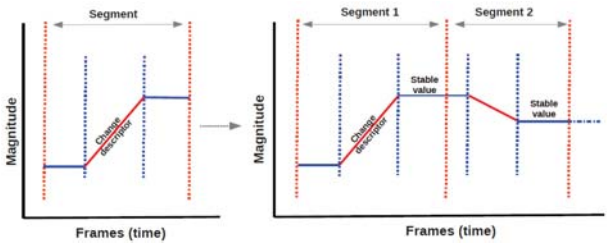


Fig. 2. *Constructing a triphone model from two separate diphone transition segments.*

constraints ($seg_{start} < T1 < T2 < seg_{end}$) prevent invalid samples from being generated.

These synthetic diphones form the building blocks for generating artificial ASR data. Construction of a complete triphone example is now simply a matter of concatenating two appropriate diphone segments for a required context, as illustrated in Figure 2. Every triphone model requires shared stable values between the channel-based piece-wise linear models of two segments. We use the implementation described in [9] for this process, and form trajectories for complete utterances from the small diphone segments.

### C. Augmenting train data triphone classes

Additional triphone examples can now be added to the training data by generating fully artificial utterances. As it is not possible to simply stitch the required triphone labels together to form an utterance directly, we define an additional 'garbage' phone that can be added whenever subsequent triphone labels do not match. Each garbage model is generated from the preceding and following diphone segment pdfs, and is discarded after training, prior to decoding.

### D. Trajectory-based feature construction

In order to generate features from trajectory models, we extend the standard Mel-frequency cepstrum coeffiecient (MFCC) feature description: similar to generating standard MFCCs, the first step is to perform a fast Fourier transform (FFT) and obtain raw filterbank outputs. We use the Hidden Markov Model Toolkit (HTK) [17] and fairly standard parameters (sampling the speech signal at a frame rate of 5ms with a set of 26 filters).

Three additional steps are required before trajectories can be estimated: the log operation, mean subtraction for every channel and ARMA filtering [14]. Trajectories are only extracted for training data. (For the test data, we rely on ARMA filtering alone to smooth features. Alternatively, two pass-recognition can be used to obtain test trajectories and possibly further improve recognition accuracy; this is not evaluated as part of the current study.)

Sampling the trajectory models estimated at this point generates a new set of frame-based features with a standard 10ms frame rate. We apply the discrete cosine transform (DCT) with a cepstral liftering coefficient of 22 (as implemented in HTK). This provides 13 cepstral features, for which the standard first and second order derivatives are taken. Lastly, cepstral mean and variance normalisation (CMVN) are applied to the complete data set.

## IV. Experimental setup

All experiments use a single-speaker corpus, specifically designed for trajectory modelling: the Afrikaans Trajectory Tracking corpus (ATT) [9]. In the next section (Section V), we first show that the trajectory models approximate the training data fairly accurately. We then analyse the mismatch between triphones in the training and test sets and experiment with different ways of creating synthetic triphones, paying specific attention to the difference between reconstructed triphones (not seen at all in the training data) and rarely seen triphones.

### A. Speech data

The ATT corpus [9] consists of about 6 000 short utterances of a single male speaker, with a 4 974 subset considered of good audio and transcription quality. From this 'clean' data set, training and test data sets were selected of 4 072 and 902 utterances respectively, as described in [9]. As this data set is still quite large, we select a random subset of 961 utterances (less than 40 minutes of speech) to construct the under-resourced training data set. A further 440 utterances were selected (also randomly), as a development set.

### B. Segmentation and test system parameters

To segment the training data for trajectory modelling, we use a standard HMM-based ASR system trained on all 4 974 utterances and perform automatic alignment of the training data. (These alignments are not used during testing.)

In all experiments to follow, similar systems are trained: a context-dependent cross-word phone recogniser with tied triphone models and 39 cepstral trajectory features (the first 13 features as defined in Section III-D, and their first and second order derivatives). These features are computed with a window size of 25ms and at a frame rate of 10ms. Tied triphone models are estimated using standard phonetic decision-tree clustering. Each triphone model has 3 emitting states with 7 Gaussian mixtures per state and a diagonal covariance matrix; semitied transforms are applied. Only insertion penalties are optimised during decoding: for all experiments this is done using the development set, with results reported on the test set.

## V. Experiments and Results

### A. Accurate trajectory-based representation

| Nr | Feature type | Channels | MSE | $\rho$ |
|---|---|---|---|---|
| 1 | Cepstral | 13 | 0.1836 | 0.9133 |
| 2 | Spectral | 26 | 0.0633 | 0.9672 |
| 3 | Spectral (ARMA 6) | 26 | 0.0177 | 0.9834 |

TABLE I

*Measuring approximation efficiency of trajectories with weighted MSE and correlation measurements.*

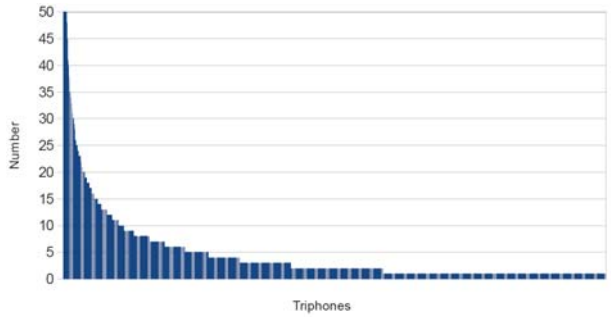


Fig. 3. *Number of triphone examples in train data (3754 triphone labels in total).*

First, we evaluate the 'goodness of approximation' of the trajectory models by measuring the variance-weighted MSE between the original feature frames and the corresponding model values (using eq. 3). These values are shown in Table I for three different sets of features, evaluated on the test data set. We also calculate the Pearson correlation coefficient ($\rho$) between trajectory estimates and actual feature values. From Table I it is clear that the two measures correlate well, that the ARMA-filtered spectral trajectories provide the best fit, and that a low MSE value ($0.0177$) and high correlation coefficient ($0.9834$) are obtained this way. The rest of the experiments only use ARMA-filtered spectral trajectories.

### B. Triphone coverage

In Table II and Figure 3 the triphone coverage of the training data is given. $3\ 754$ unique triphones are observed. On average, each is observed 5 times, with only about a $\frac{1}{4}$ of labels occurring more than 5 times. Good overlap ($2\ 608$ labels) exists between the training and test data sets. Of the $1\ 045$ labels not seen in the training set, $929$ can be reconstructed from diphones (that is, both required diphones are observed in the training data).

| Category | Triphone count |
|---|---|
| Train data | 3 754 |
| Test data | 3 653 |
| Test data seen | 2 608 |
| Test data unseen | 1 045 |
| Test data unseen: Diphone constructable | 929 |
| Test data unseen: Not constructable from diphones | 116 |

TABLE II

*Triphone overlap between test and train data sets and that the number re-created from diphone transitions.*

Figure 3 clearly shows how rarely seen triphones can also be exploited. We experiment with four pools of these labels: (1) $< 2$ examples with $692$ labels, (2) $< 3$ examples - $1127$ labels, (3) $< 5$ examples - $1616$ and (4) $< 10$ examples where $2189$ labels form part of the rare category.

### C. Baseline

Evaluating the phone recognition accuracy of the test data, leads to the baseline result ('Control') in Table III. Phone accuracies are reported for systems with and without semitied transforms.

| System | #Cons | #Rare | ACC | ACC (semitied) |
|---|---|---|---|---|
| Control | - | - | 78.31 | 78.65 |
| Reconstruct 1 | 1 | - | 78.66 | 79.54 |
| Reconstruct 2 | 2 | - | 79.33 | 79.87 |
| Reconstruct 3 | 3 | - | 78.76 | 79.40 |
| Reconstruct 4 | 5 | - | 78.26 | 79.16 |
| Reconstruct 5 | 10 | - | 76.54 | 79.08 |
| Rare 1 | - | 2 | 79.26 | 79.70 |
| Rare 2 | - | 3 | 79.95 | 80.04 |
| Rare 3 | - | 5 | 79.56 | 80.39 |
| Rare 4 | - | 10 | 77.34 | 79.31 |
| Combined 1 | 2 | 2 | 78.47 | 80.55 |
| Combined 2 | 1 | 3 | 79.48 | 80.82 |
| Combined 3 | 2 | 3 | 79.46 | 79.99 |
| Combined 4 | 3 | 3 | 79.39 | 80.40 |
| Combined 5 | 3 | 5 | 78.89 | 80.30 |
| Combined 6 | 5 | 5 | 78.53 | 79.54 |
| Combined 7 | 10 | 10 | 75.96 | 78.33 |

TABLE III

*Phone recognition results when adding synthetic triphone examples to rare and unseen triphones in the training data.*

### D. Reconstructed triphones

We experiment with five acoustic models (Reconstruct 1 - 5), reconstructing unseen triphones; these only differ with regard to the number (#Cons) of reconstructed synthetic triphone examples generated per label (Table III). From these results, adding unseen triphones to the training data does improve phone recognition accuracy. Interestingly, only adding a few samples works well: adding too many examples ($5$ or $10$) does not provide better accuracy.

### E. Rare triphones

Since at least one example of real training data is seen for the triphone identities in the rare triphone category, a smaller number of rarely seen triphone examples needs to be added to achieve the same set number of examples per triphone label. The results in Table III prove that the seen examples are not adequate: system accuracy improves significantly when adding synthetic triphone examples.

For the next four acoustic models, we steadily increase the number of synthetic triphones in the rare triphone category. In the Rare 1 experiment, at least 2 examples of all triphones are included (after additional triphones have been generated). Similarly, for systems Rare 2-4, at least 3, 5 and 10 examples are included in the training data. (See Table II for the number of triphones in each class.)

Again, adding 10 examples does not provide the best improvement ($79.31\%$). The Rare 3 model provides a better result than that obtained for unseen triphones (Reconstruct 2 model).

### F. Under-resourced triphones

Given the results above, the next question is whether generating synthetic triphone examples for both the rarely seen and unseen triphone categories would contribute to

gains in phone recognition accuracy. In an attempt to do so, we combine the synthetic triphone example sets.

Forcing 10 triphone examples remains too many. The accuracy of 78.33% that the Combined 7 system achieves (Table III) show no gain over the baseline. Lowering the number of synthetic triphone examples to 5, significantly improves accuracy, but still does not outperform the previous Rare triphone experiments. In fact, the Combined 4 system results match these, but now for the system where we force a number of 3 examples of the under-resourced triphone label classes.

Since the reconstructed and rare triphone categories behave differently, as a last refinement, we also test what happens when different numbers of examples are added from each triphone class. We obtain the best results when a single triphone example is added for each of the 926 labels of the unseen category and at least 3 seen examples for the 1 127 labels of the rare category is forced. These results are tuned to the specific data set considered here: we do not propose it as a general strategy for adding synthetic triphones. Rather, we find it interesting that recognition accuracy can be improved using a fairly crude strategy for generating synthetic triphones.

## VI. Discussion

When developing ASR systems in resource-constrained environments, many triphones are never seen during training. We found that it is possible to reconstruct unseen triphones from smaller contextual units, and that this can improve recognition accuracy of a standard tied-state baseline ASR system. Although speech synthesis techniques could also be used to generate unseen triphone units, we found that the trajectory models we use are able to represent training data surprisingly well. Our technique leads to a natural process for creating artificial utterances, containing repeated sequences of synthetically created samples of both unseen and rare triphones. Randomly selecting the training data set from the phonetically balanced ATT corpus [9], we can expect the distributions of unseen and rare triphones to remain comparable for new speakers of the same language.

With the current approach, the number of samples that can be added is still quite limited. Further refinements to the model are foreseen, especially with regard to the current sampling process, which is fairly crude. Trajectory optimisation and exploring the relationship of the new trajectories and the MSE or correlation measures in Table I may lead to improved results. Similarly, we report on improved results for semitied transforms, but the effect using training features with a high degree of feature smoothness on this technique remains to be investigated.

Our goal is to first analyse and understand the current restricted environment (a single speaker, generating triphones from diphones) in more depth, before considering the extent to which the current approach can generalise to larger contexts (quinphones from triphones) and finally, cross-speaker data augmentation. This is aligned with recent data augmentation approaches that have begun to address the cross-speaker problem. For example, [18] attempts to find a stochastic feature mapping (SFM) to statistically convert features between two speakers. The extent to which these approaches (generating additional synthetic contexts for a single speaker, and generating additional synthetic speakers) are complementary, remains an open question.

## References


[1] K.-F. Lee, "Large-vocabulary speaker-independent continuous speech recognition: The sphinx system," Ph.D. dissertation, Carnegie Mellon University, 1988.

[2] H. Chang and J. Glass, "Multi-level context-dependent acoustic modeling for automatic speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, Waikoloa, HI, December 2011, pp. 89–94.

[3] A. Ragni, K. Knill, S. Rath, and M. Gales, "Data augmentation for low resource languages," in *Proceedings of Interspeech*, Singapore, September 2014, pp. 810–814.

[4] H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C.-H. Lee, "A study on multilingual acoustic modeling for large vocabulary ASR," in *Acoustics, Speech and Signal Processing (ICASSP), 2009 IEEE International Conference on*, Taipei, April 2009, pp. 4333–4336.

[5] N. Jaitly and G. E. Hinton, "Vocal tract length perturbation (VTLP) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language Processing*, 2013.

[6] N. Kanda, R. Takeda, and Y. Obuchi, "Elastic spectral distortion for low resource speech recognition with deep neural networks," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 309–314.

[7] J. A. C. Badenhorst, M. H. Davel, and E. Barnard, "Analysing co-articulation using frame-based feature trajectories," in *Proceedings of PRASA*, Stellenbosch, South Africa, November 2010, pp. 13–18.

[8] J. Badenhorst, M. Davel, and E. Barnard, "Trajectory behaviour at different phonemic context sizes," in *Proceedings of PRASA*, Vanderbijlpark, South Africa, November 2011, pp. 1–6.

[9] J. Badenhorst, M. Davel, and E. Barnard, "Improved transition models for cepstral trajectories," in *Proceedings of PRASA*, Pretoria, South Africa, November 2012, pp. 157–164.

[10] V. Digalakis, "Segment-based stochastic models of spectral dynamics for continuous speech recognition," Ph.D. dissertation, Boston University, 1992.

[11] W. Holmes and M. J. Russell, "Probabilistic-trajectory segmental HMMs," *Computer Speech and Language*, vol. 13, no. 1, pp. 3–37, January 1999. [Online]. Available: http://www.idealibrary.com

[12] L. Zhang and S. Renals, "Phone recognition analysis for trajectory HMM." in *Proc. Interspeech*, 2006.

[13] D. Yu, L. Deng, and A. Acero, "A lattice search technique for a long-contextual-span hidden trajectory model of speech," *Speech Communication*, vol. 48, no. 9, pp. 1214–1226, September 2006.

[14] C. P. Chen and J. A. Bilmes, "MVA processing of speech features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 257–270, January 2007.

[15] M. U. X. Lu, S. Matsuda and S. Nakamura, "Temporal contrast normalization and edge-preserved smoothing of temporal modulation structures of speech for robust speech recognition," *Speech Communication*, vol. 52, no. 1, pp. 1–11, January 2010.

[16] S. C. X. Xiao and H. Li, "Normalization of the speech modulation spectra for robust speech recognition," *IEEE Transactions on Computer Speech and Language*, vol. 16, no. 8, pp. 1662–1674, November 2008.

[17] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book*. http://htk.eng.cam.ac.uk/: Cambridge University Engineering Department, 2005.

[18] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference*. IEEE, 2014, pp. 5582–5586.