Phonetic Segmentation of Speech using STEP and t-SNE

Adriana Stan[1], Cassia Valentini-Botinhao[2], Mircea Giurgiu[1], Simon King[2]

[1]Communications Department
Technical University of Cluj-Napoca, Romania
{Adriana.Stan, Mircea.Giurgiu}@com.utcluj.ro

[2]Centre for Speech Technology Research
University of Edinburgh, United Kingdom
cvbotinh@inf.ed.ac.uk, Simon.King@ed.ac.uk

***Abstract*—This paper introduces a first attempt to perform phoneme-level segmentation of speech based on a perceptual representation -- the Spectro Temporal Excitation Pattern (STEP) -- and a dimensionality reduction technique -- the t-Distributed Stochastic Neighbour Embedding (t-SNE). The method searches for the true phonetic boundaries in the vicinity of those produced by an HMM-based segmentation. It looks for perceptually-salient spectral changes which occur at these phonetic transitions, and exploits t-SNE's ability to capture both local and global structure of the data. The method is intended to be used in any language and it is therefore not tailored to any particular dataset or language. Results show that this simple approach improves segmentation accuracy of unvoiced phonemes by 4% within a 5 ms margin, and 5% at a 10 ms margin. For the voiced phonemes, however, accuracy drops slightly.**

Keywords— phonetic segmentation; STEP; t-SNE; HMM acoustic model; k-Means.

I. INTRODUCTION

The basic unit of speech production and perception is widely accepted as being the phoneme. Knowing the exact location of the phones and their boundaries is essential for some applications using machine learning algorithms to process speech data. However this is not a trivial task, as even in manual segmentation the inter-labeller agreement as to where the phone boundaries should be placed is around 93% [1] within a 20 ms margin.

Accurate phonetic segmentation might not be as important in all speech-enabled applications, such as those based on stochastic training algorithms like HMM-based speech recognition or synthesis. But it is essential in newly developed fields, such as automatic lip-syncing [2], where even a few milliseconds deviation from the boundary might lead to unnatural output.

Previous studies on phonetic segmentation have considered various methods and feature sets. The most common method for this task is forced alignment with HMM acoustic models [3], with slight variations in the training [1, 4] and decoding [5, 6] methods, or the feature sets used [7, 8]. One drawback of this category of methods is the need for a phonetic transcription of the speech data. When the phonetic string is not known, variations in the spectral and temporal features, or representations of speech called landmarks, are one basis for estimating phone boundaries [9, 10, 11]. More recent studies have also exploited the capabilities of Deep Belief Networks to estimate the posterior probabilities of phone categories and to then assign boundaries to frames where there is uncertainty in assigning a phone class [12].

What differentiates our work from these previous approaches is that we try to avoid tailoring our method towards a particular language or speech database. Therefore, although our results do not achieve the same accuracy as the best methods evaluated on the same dataset, namely TIMIT [13], the method can be easily extended to a variety of other speech resources. The core of our method relies on the fact that spectral discontinuities are in most cases a good indicator of a phoneme boundary, even when these transitions are smooth, such as in the case of diphthongs. In this respect, our method would fall into the landmark category of phonetic alignment, but we also use the forced alignment as a reference to limit the search space for these landmarks. Computing distances between consecutive frames in high-dimensional acoustic feature space is subject to the so-called "dimensionality curse", in which the number of equally distanced data points grows exponentially with the number of dimensions [14]. Therefore, we first reduce the dimensionality of the acoustic space to 2 or 3 dimensions. One other aspect to be noted is the fact that variations in the parametrisation vectors do not necessarily correspond to salient acoustic changes. Therefore, a perceptual representation of the speech signal is used. Both the representation and the dimensionality reduction method are presented in Section II, and the evaluation of their performance is presented in Section III.

II. PROPOSED PHONETIC SEGMENTATION

The proposed method uses the Spectro Temporal Excitation Pattern (STEP) parametrisation, reduced in dimension by t-Distributed Stochastic Neighbour Embedding (t-SNE). This section describes the details of both STEP and t-SNE, followed by an overview of the processing flow for the phonetic segmentation.

A. STEP

The STEP representation was originally proposed in the context of the Glimpse model for speech perception in noise [15]. The model is motivated by the ability of humans to obtain information from time-frequency regions where speech is not masked by noise and therefore less distorted. The Glimpse Proportion (GP) measure proposed in [16] is based on this

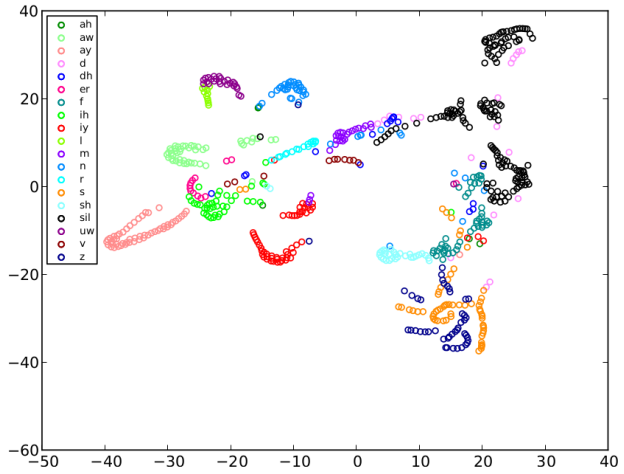

Fig. 1. t-SNE 2D mapping of the utterance "*The reasons for this dive seemed foolish now",* and the Praat visualisation of the waveform, spectrogram, and manual annotation. The axes represent the 2 t-SNE dimensions.

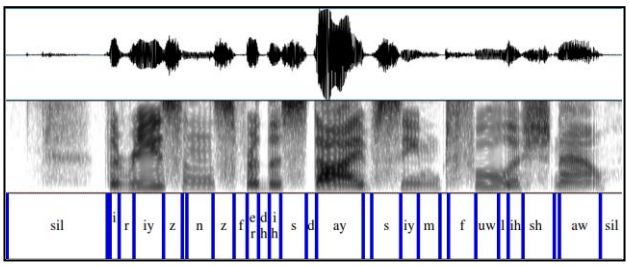

Fig. 2. Praat visualisation of the waveform, spectrogram, and manual annotation of the utterance "*The reasons for this dive seemed foolish now".*

concept: in a noisy environment, humans focus their auditory attention on 'glimpses' of speech that are not masked by noise. To detect such glimpses, the STEP representations of speech and noise are compared. The GP correlates well with subjective scores for intelligibility of both natural [16] and synthetic speech [17] in a variety of noises.

To represent a signal in terms of STEP we first decompose its waveform into different frequency channels using a Gammatone filterbank whose central frequencies are linearly spaced on the Equivalent Rectangular Bandwidth (ERB) scale [18]. For each channel, the temporal envelope is extracted with an absolute value operation, smoothed with a low pass filter and then averaged across limited time intervals.

B. t-SNE

t-Distributed Stochastic Neighbor Embedding (t-SNE) [19, 20] is a dimensionality reduction technique targeted mostly at high-dimensional data visualization. As opposed to other dimensionality reduction algorithms, t-SNE is capable of capturing both local and global structure of the data. This allows for visualizing similar data points in local regions, or globally-emerging clusters.

t-SNE is a variation of Stochastic Neighbor Embedding (SNE) [21] which converts pairwise Euclidean distances in N-

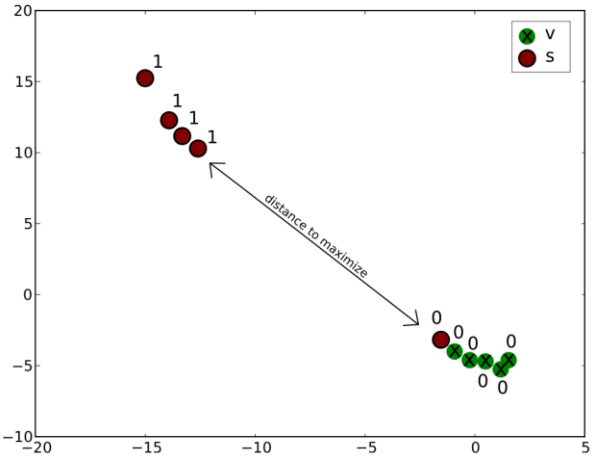

Fig. 3. Example of a k-Means clustering result in the vicinity of a forced-alignment boundary. Different colors represent different phone identities as determined by the forced alignment, while the numbers represent the clusters to which each point has been assigned. The axes represent the 2 t-SNE dimensions.

dimensional space, to joint probability distributions. Given a set of N-dimensional data points $X = \{x_1, x_2, \ldots x_n\}$, the joint probability distributions P are computed as:

$$p_{ij} = \frac{e^{-d_{ij}^2/\sigma}}{\sum_k \sum_{l \neq k} e^{-d_{kl}^2/\sigma}} \quad (1)$$

where $d_{ij} = \|x_i - x_j\|^2$ is the N-dimensional norm, σ is the variance of the Gaussian distribution centered on datapoint x_i, and $p_{ii}=0$.

The low-dimensional mapping obtained by t-SNE, $Y = \{y_1, y_2, \ldots y_n\}$ uses a Student-t distribution with a single degree of freedom to model the similarity between two data points:

$$q_{ij} = \frac{(1+d_{ij})^{-1}}{\sum_k \sum_{l \neq k} (1+d_{kl})^{-1}} \quad (2)$$

where $d_{ij} = \|y_i - y_j\|^2$ is the low-dimensional norm, and $q_{ii}=0$. The obtained mapping minimizes the Kullback-Leibler divergence with respect to the high-dimensional distribution, using a gradient descent method.

C. Phonetic Segmentation with STEP and t-SNE

A baseline forced-alignment using HMM models provides a good set of reference points near which the true phone boundaries should be located. The proposed method searches for the correct phonetic boundary in the vicinity of these reference points.

Each utterance is processed individually by first extracting the STEP features as detailed in Section II-A. The STEP

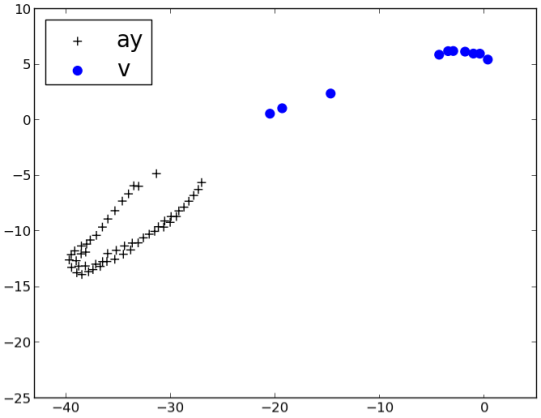

Fig. 4. Example of a 2D t-SNE representation at a voiced-unvoiced boundary. The axes represent the 2 t-SNE dimensions.

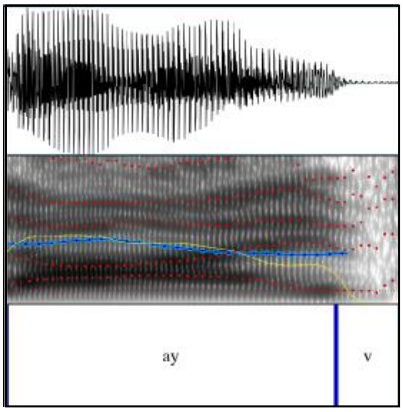

Fig. 5. Praat visualisation of the waveform, spectrogram and manual annotation at a voiced-unvoiced boundary type.

features are then scaled by subtracting the mean and scaling to unit variance. This ensures a uniform distribution of the features across the utterances, removing speaker characteristics to some extent.

Afterwards, the t-SNE dimensionality reduction method is applied over the STEP features. As t-SNE is stochastic, the results from successive runs may vary slightly. To mitigate, we run t-SNE 5 times for each utterance. The best run is selected by using the k-Means clustering algorithm. k-Means is applied to a subset of frames located around the baseline forced-alignment boundary, starting from the previous t-SNE+STEP boundary[1] to the next forced alignment segment boundary.

The best run is considered to be the one in which the maximum distance between consecutive frames assigned to separate clusters is obtained (see Figure 3). This best run may vary from boundary to boundary, and therefore independent representations are maintained for each of them.

Figures 1 and 4 show an example of a 2D t-SNE representation at the utterance level, as well as around a phone boundary in the data. In Figure 1, it can be observed that the different phone identities tend to cluster together at a global level, while Figure 4 shows the spectral changes that can occur within the same phonetic segment.

This dimensionality reduction enables computation of the Euclidean distance between successive frames, without the curse of dimensionality. Therefore the maximum distance between consecutive frames in the neighborhood of the forced alignment window is assigned to be the new phone boundary (i.e. /v/). Of course, there is a possibility that the window centered on the initial alignment does not include the true boundary, and therefore our method cannot correct it.

[1] If this is not the first boundary in the utterance.

TABLE 1. Accuracy of the forced-alignment phone boundary assignment for the MFCC and MFCC+STEP acoustic models at 5, 10, 20 and 50 ms threshold.

System	Accuracy [%]			
	5ms	**10ms**	**20ms**	**50ms**
MFCC	39.68	56.76	83.34	92.33
STEP	37.12	55.22	80.00	89.76
MFCC+STEP	42.93	62.53	84.29	94.17

III. EVALUATION

A. Data

In line with previous studies on phone-level segmentation, we selected the TIMIT dataset [13, 22] as a reference for our evaluation. It comprises approximately 5.5 hours of recordings of phonetically-balanced prompted speech uttered by 630 speakers in 8 major dialects of American English. The dataset is split into a training and a testing set. Excluding the dialect calibration tests (i.e., the *sa* sentences), the training set contains 3696 utterances from 462 speakers, approximately 3.14 hours of data; and the test set contains 1344 utterances from 168 speakers, approximately 1.5 hours of data. The dataset is sampled at 16 kHz with a 16 bit resolution. The 61 phones used in the annotation of TIMIT were mapped to the CMU Pronouncing Dictionary[2], resulting a set of 40 phones. The silence segment boundaries were excluded from the evaluation, yielding a total of 40,516 segment boundaries in the test set.

B. Baseline Alignment System

HMM acoustic model-based alignment is widely used in both speech synthesis and recognition systems, and considered to be accurate enough not to affect the performance of the resulting system. However, when compared to manual segmentation, its accuracy within a 20 ms threshold is somewhere between 80-90%, depending on the features or the training method used [1]. The best results obtained using HMMs for phonetic segmentation on the TIMIT database are reported in [4] and establish an alignment accuracy of 96.7% within a 20 ms margin. However these models are tuned for the TIMIT database itself, and when using other datasets the models would have to be tuned once again, which is computationally complex and resource consuming.

[2] Available online: http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

TABLE 2. Segmentation results for the baseline system, t-SNE with low-dimensional spaces of 2 and 3 features using STEP and MFCC.

System	Accuracy [%]			
	5ms	10ms	20ms	50ms
Baseline	42.93	62.53	82.29	94.17
t-SNE 2D+STEP	41.34	59.52	77.73	89.54
t-SNE 3D+STEP	41.89	60.49	79.90	91.78
t-SNE 2D+MFCC	38.12	57.00	76.12	88.43
t-SNE 3D+MFCC	39.72	57.12	77.09	88.20

TABLE 3. Accuracy of the segmentation for the voiced and unvoiced phonetic categories. t-SNE uses STEP as the initial feature space.

Voiced phones				
System	Accuracy [%]			
	5ms	10ms	20ms	50ms
Baseline	39.71	59.54	80.59	92.12
t-SNE 2D	36.88	54.50	73.20	86.16
t-SNE 3D	38.31	56.28	76.28	89.09

Unvoiced phones				
System	Accuracy [%]			
	5ms	10ms	20ms	50ms
Baseline	41.37	58.79	78.11	89.44
t-SNE 2D	45.97	63.57	79.23	87.85
t-SNE 3D	44.81	62.94	79.51	88.58

We therefore start from a baseline forced-alignment of the TIMIT training set speech data with HMM acoustic models. Three separate models were built: one using standard 13 MFCCs with energy, delta and delta-deltas; one using 34 STEP with energy, delta and delta-deltas; and one using both the MFCC and STEP representation: 34 STEP coefficients extracted at each 5 ms frame and 13 MFCCs. Energy, delta and delta-deltas were appended to the fused features. All acoustic models used a 5 state, left-to-right, context-independent HMM for each phone.

The manual segmentation was not used for model training, but instead an iterative alignment and training procedure was performed. So, the results of the method will generalise to datasets with no manually annotated training set.[3]

To extract STEP, we used 34 Gammatone filters whose central frequencies covered the range of 100-7500 Hz and the temporal integration time for the smoothing filter was of τ (8 ms). The original STEP representation was calculated with non-overlapping time frames using rectangular windows, but for a smoother trajectory we calculate it using a F0 adaptive window as in [24] of 40 ms length and 5 ms shift.

C. Results

All the results reported in this section are measured against the manual segmentation of the test set data from the TIMIT database. The threshold, given in milliseconds, represents the allowed deviation of an estimated phone boundary from the manual annotation. The values selected for the threshold are 5, 10, 20 and 50 ms, as used in previous studies of phonetic segmentation. The accuracy is computed as the number of estimated phoneme boundaries lying within the threshold, divided by the total number of boundaries.

[3] In [23] the authors report only 2% difference in accuracy at 20 ms threshold for their highly tuned models: the iterative training procedure is likely to produce similar results to the fully supervised one.

TABLE 4. Accuracy of the segmentation for the different boundary types. t-SNE uses STEP as the initial feature space.

Unvoiced-Unvoiced				
System	Accuracy [%]			
	5ms	10ms	20ms	50ms
Baseline	22.27	35.03	53.58	60.86
t-SNE 2D	31.70	43.76	54.79	60.26
t-SNE 3D	30.76	42.52	54.46	60.65

Voiced-Unvoiced				
System	Accuracy [%]			
	5ms	10ms	20ms	50ms
Baseline	41.85	59.34	79.25	87.89
t-SNE 2D	46.77	64.27	79.02	86.69
t-SNE 3D	45.69	63.55	79.41	87.34

Unvoiced-Voiced				
System	Accuracy [%]			
	5ms	10ms	20ms	50ms
Baseline	42.32	63.36	81.76	87.84
t-SNE 2D	39.66	58.93	79.40	86.90
t-SNE 3D	38.80	58.40	79.52	87.06

Voiced-Voiced				
System	Accuracy [%]			
	5ms	10ms	20ms	50ms
Baseline	36.99	53.91	71.88	87.61
t-SNE 2D	30.54	42.50	61.28	78.06
t-SNE 3D	33.55	48.43	66.11	82.46

1) Comparing baselines

A first evaluation of our method concerns the features used for the HMM-based acoustic model training. The results of using either MFCC, STEP or a fusion of MFCCs and STEPs are presented in Table 1. The acoustic models which use only STEPs perform worse than the ones with MFCCs, and this might be caused by the fact that although STEPs are better suited for perceptual representations, their smoothed trajectories limit the method. However, the fused MFCC+STEP models were marginally better at the 20 and 50 ms threshold, and significantly better at the lower threshold of 5 and 10 ms. Given this improvement in fine-grained segmentation, we selected the MFCC+STEP acoustic models as a baseline alignment for the t-SNE and STEP method.

2) Comparing methods

The t-SNE is an algorithm designed for data visualization in 2D or 3D spaces. But, given its inherent properties, it also provides a good dimensionality reduction and clustering method. As there is no mathematical way to determine which of the two low-dimensional spaces is better suited for this scenario, we compared them on our test data, and also against the baseline forced-alignment. We also examined the use of

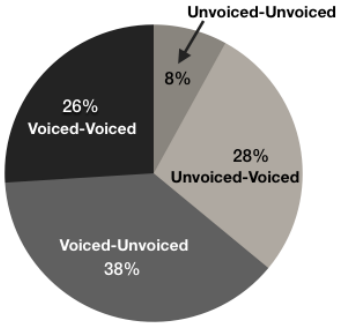

Fig. 6. Speech corpus division into phonetic boundary categories.

MFCCs for the phonetic segmentation. The results are shown in Table 2. At a first glance, t-SNE appears to increase segmentation errors, and when used on top of MFCCs, it performs the poorest. However, by examining the 2D t-SNE representations in Figures 1 and 4, t-SNE does seem to be capturing the spectral changes which occur at the phone boundaries.

Motivated by this, we investigated the broad phonetic categories separately. There are 16 unvoiced phones and 23 voiced phones in the CMU phoneset, and the data is split into 35% unvoiced[4] and 65% voiced phones. The results for each category are presented in Table 3. It appears that for the unvoiced phonemes, the performance of t-SNE+STEP is above that of the baseline, especially at the finer grained margins of 5 and 10 ms.

Yet it is also important to look at the boundary types and analyse their individual accuracies. We split them into four categories based on the neighbouring phoneme identities: *unvoiced-unvoiced*, unvoiced-*voiced*, *voiced-unvoiced* and *voiced-voiced*. The percentage of each boundary type within the test set are shown in Figure 6, while the alignment results for each of the categories are shown in Table 4.[5] There is once again a distinction between the performance of our method and the baseline alignment for different boundary types. For the unvoiced-unvoiced and voiced-unvoiced types, especially at finer grained thresholds (i.e., 5 ms and 10 ms), our method outperforms the baseline alignment; in the other two cases, the baseline achieves a smaller error.

3) Discussion

As seen in the previous Section, our proposed method can identify the starting point of unvoiced phonemes with better accuracy than the baseline forced-alignment system, but it does not do such a good job at identifying the starting points of voiced phonemes. One explanation would be that there is less abrupt spectral change at the start of this phonetic category. A solution would be to adapt the distance measure between consecutive frames according to the boundary type. The fact that the 3D representation achieves slightly better results than the 2D for the voiced-voiced boundary type could also mean that such reduced dimension spaces cannot entirely capture these transitions, and that perhaps a combination of t-SNE with some other feature reduction algorithm would be beneficial.

There is also a ceiling effect to be noticed in the case of unvoiced-unvoiced boundaries, where the baseline alignment even at a 50 ms threshold has an accuracy of only 60.86%. This means that the search for the true boundary using the t-SNE+STEP method was not executed within a correct window, thus limiting its potential accuracy.

[4]Excluding silence segments.

[5]In Table 3 the segmentation results are for the case when the voiced or unvoiced phoneme is at the right hand side of the boundary. This means that it only evaluates the starting point of the phoneme.

IV. Conclusions

This paper introduced an approach to phonetic segmentation using a perceptual feature set, STEP and a dimensionality reduction method, t-SNE. These first results show that the combination of these two methods yields better results in the detection of unvoiced phone boundaries, although voiced phone boundaries are more accurately located found using the baseline method. However, the fact that the method does not rely on any language or database specificities makes it feasible for other speech resources as well. As future work, the most important development would be in the use of alternative or auxiliary features for each phonetic category, and perhaps an additional distance measure or spectral change estimator in the t-SNE feature space.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement N° 287678 (Simple4All), PN-II-PT-PCCA-2013-4 N° 6/2014 (SWARA) and the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

References

[1] J.-P. Hosom, “Speaker-independent phoneme alignment using transition-dependent states,” Speech Communication., vol. 51, no. 4, pp. 352–368, April. 2009

[2] G. Hofer and K. Richmond, "Comparison of HMM and TMDN methods for lip synchronisation," in Proc. Interspeech, Makuhari, Japan, pp. 454-457, September 2010.

[3] F. Brugnara, D. Falavigna, and M. Omologo, “Automatic segmentation and labeling of speech based on Hidden Markov Models.” Speech Communication, vol. 12, no. 4, pp. 357–370, 1993.

[4] A. Stolcke, N. Ryant, V. Mitra, W. Wang, and M. Liberman, “Highly Accurate Phonetic Segmentation Using Boundary Correction Models and System Fusion,” in Proc. ICASSP, May 2014.

[5] D. Toledano, L. Gomez, and L. Grande, “Automatic phonetic segmentation,” IEEE Trans. on Speech and Audio Processing, vol. 11, no. 6, pp. 617–625, November 2003.

[6] I. Mporas, T. Ganchev, and N. Fakotakis, “Phonetic segmentation using multiple speech features,” International Journal of Speech Technology, vol. 11, no. 2, pp. 73–85, 2008.

[7] M. Karnjanadecha and S. A. Zahorian, “Toward an optimum feature set and HMM model parameters for automatic phonetic alignment of spontaneous speech.” in Proc. Interspeech, September 2012.

[8] O. Kalinli, “Automatic phoneme segmentation using auditory attention features.” in Proc. Interspeech, September 2012.

[9] V. Peddinti and K. Prahallad, “Exploiting phone-class specific landmarks for refinement of segment boundaries in TTS databases.” in Proc. Interspeech, pp. 429–432, August 2011.

[10] S. Dusan and L. R. Rabiner, “On the relation between maximum spectral transition positions and phone boundaries.” in Proc. Interspeech, September 2006.

[11] R. Prasad and B. Yegnanarayana, “Acoustic segmentation of speech using zero time liftering (ZTL),” in Proc. Interspeech, pp. 2292–2296, August 2013.

[12] O. Kalinli, “Combination of auditory attention features with phone posteriors for better automatic phoneme segmentation.” In Proc. Interspeech, pp. 2302–2305, August 2013.

[13] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus CDROM,” 1993.

[14] C. Sammut and G. I. Webb, Eds., Encyclopedia of Machine Learning. Springer, 2010.

[15] M. Cooke, “Glimpsing speech,” Journal of Phonetics, vol. 31, pp. 579 – 584, 2003.

[16] M. Cooke, “A glimpsing model of speech perception in noise,” J.Acoust. Soc. Am., vol. 119, no. 3, pp. 1562–1573, 2006.

[17] C. Valentini-Botinhao, J. Yamagishi, and S. King, “Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?” in Proc. Interspeech, pp. 1837 – 1840, August 2011.

[18] B. C. J. Moore and B. R. Glasberg, “A revision of Zwicker’s loudness model,” Acta Acustica, vol. 82, pp. 335–345, 1996.

[19] L. van der Maaten and G. E. Hinton, “Visualizing High-Dimensional Data Using t-SNE,” Journal of Machine Learning Research, vol. 9, pp. 2579–2605, 2008.

[20] L. van der Maaten and G. Hinton, “Visualizing non-metric similarities in multiple maps,” Machine Learning, vol. 87, no. 1, pp. 33–55, 2012.

[21] G. Hinton and S. Roweis, “Stochastic neighbor embedding,” Advances in neural information processing systems, vol. 15, pp. 833–840, 2003.

[22] C. Lopes and F. Perdigao, “Phone Recognition on the TIMIT Database,” Speech Technologies, 2011.

[23] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, “Automatic phonetic segmentation using boundary models,” in Proc. Interspeech, Lyon, France, Aug. 2013, pp. 2306–2310.

[24] H. Kawahara, I. Masuda-Katsuse, and A. Cheveign, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds,” Speech Communication, vol. 27, pp 187-207, 1999.