Phonetic Segmentation using STEP and t-SNE

Adriana Stan, Cassia Valentini-Botinhao,
Mircea Giurgiu, Simon King

SpeD 2015
Bucharest, Romania
October 14-17, 2015

Natural Speech Technology

Outline

- Brief review of phone-level segmentation
- Proposed method using STEP and t-SNE
- Evaluation and results
- Conclusions and future work

Phone-level segmentation of speech

Phone level segmentation of speech

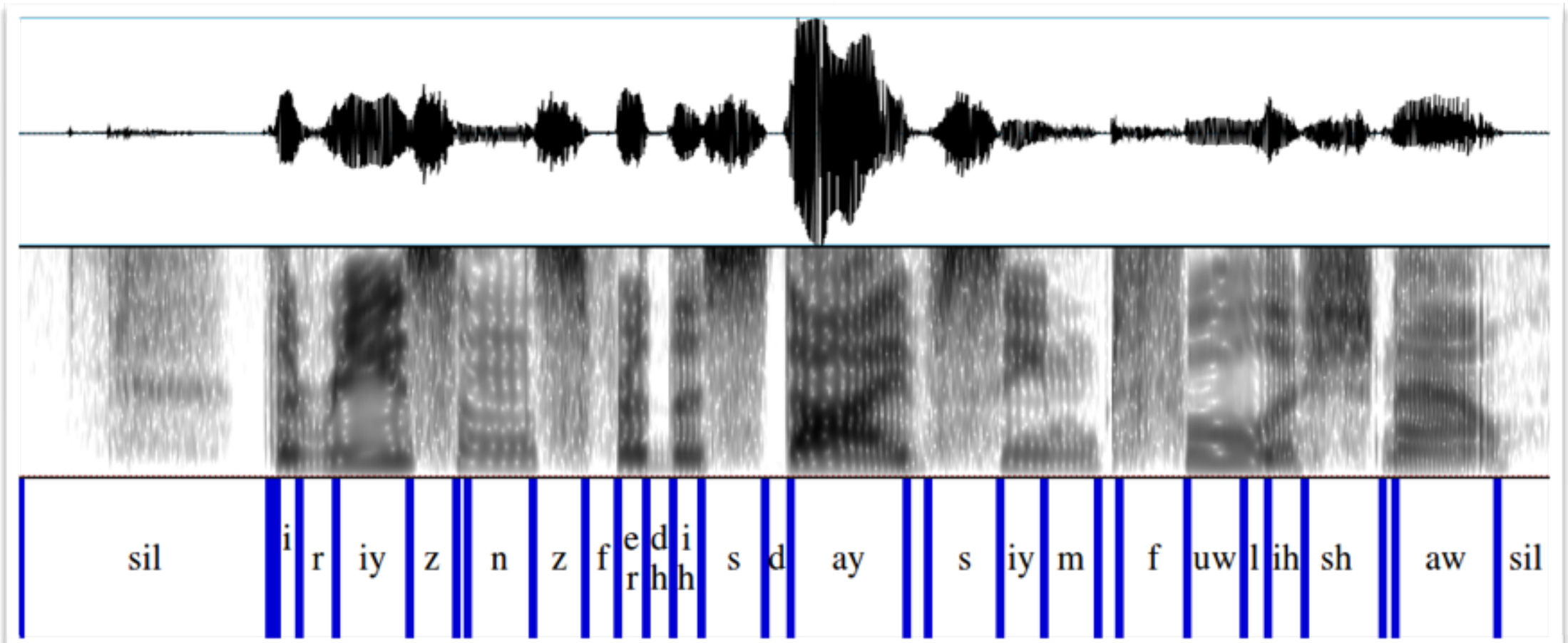


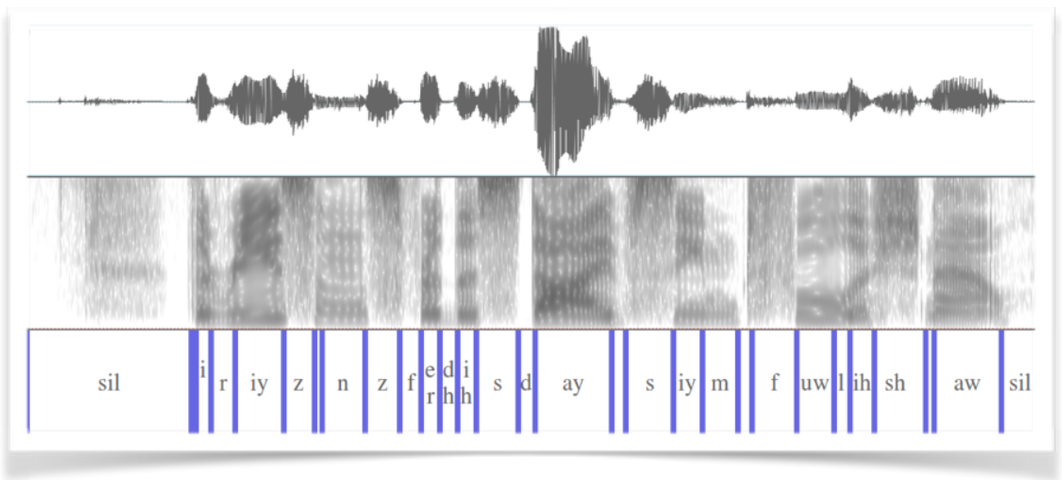
Fig. 1. Praat visualisation of a phonetic segmentation of the utterance “The reasons for this dive seemed foolish now”.

Phone-level segmentation of speech

Methods

- HMM-based acoustic modelling [1,2,3,4,5]
- landmark detection [6,7,8]
- deep belief networks [9]

Proposed phonetic segmentation



Proposed method

Main ideas

- Avoid tailoring the method towards a particular language or dataset;
- Spectral discontinuities are in most cases a good indicator of a phoneme boundary;
- The “curse of dimensionality” could be avoided by using a good dimensionality reduction technique;
- Use a perceptually relevant acoustic parametrisation.

Proposed method

Solutions

- HMM-based forced alignment can provide a good reference for phonetic boundaries;
- t-Distributed Stochastic Neighbour Embedding dimensionality reduction;
- Spectro-Temporal Excitation Pattern parametrisation.

Proposed method

t-SNE

- Merck Viz Challenge winning dimensionality reduction technique;
- converts pairwise Euclidean distances in N-dimensional spaces into joint probability distributions
- in low-dimensional space the similarity between two data points is modelled by a Student-t distribution;
- the mapping minimises the Kullback-Leibler divergence with respect to the high-dimensional distribution, using a gradient descent method.

Proposed method

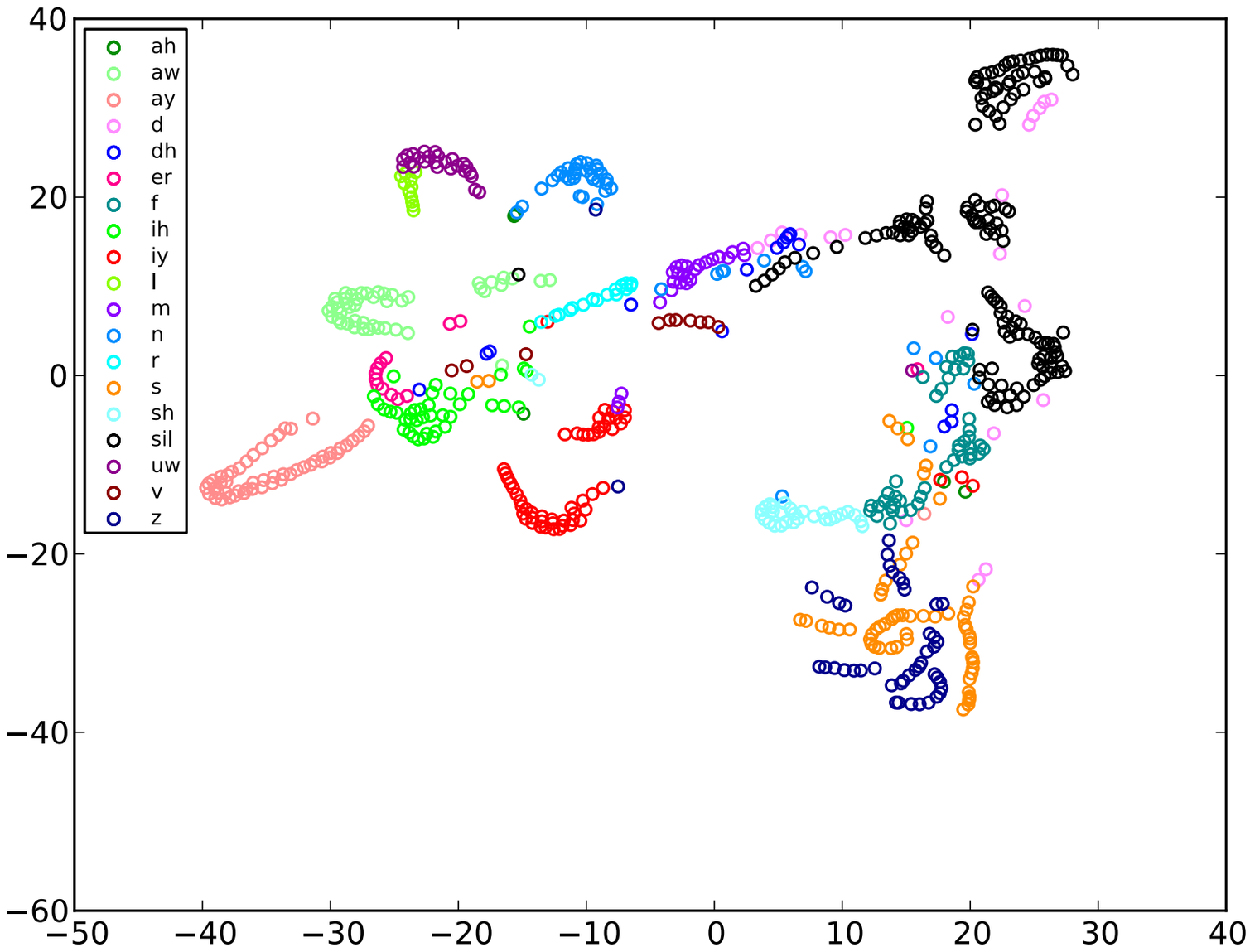


Fig. 2. t-SNE 2D representation of the utterance “The reasons for this dive seemed foolish now”.

Proposed method

STEP parametrisation

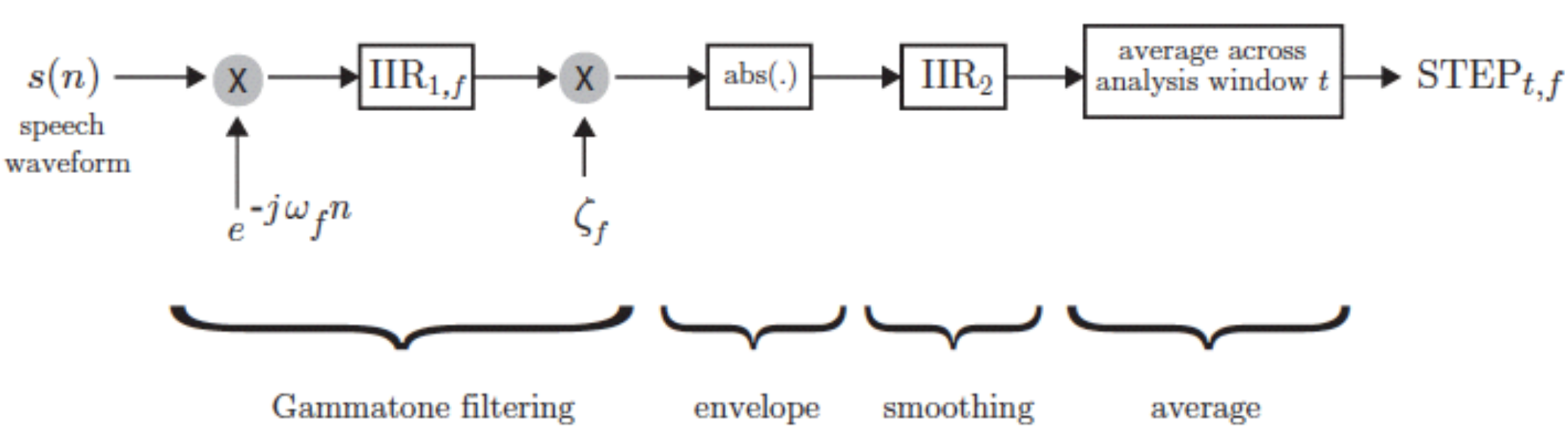


Fig. 3. STEP calculation

Proposed method

Utterance

Proposed method

Utterance

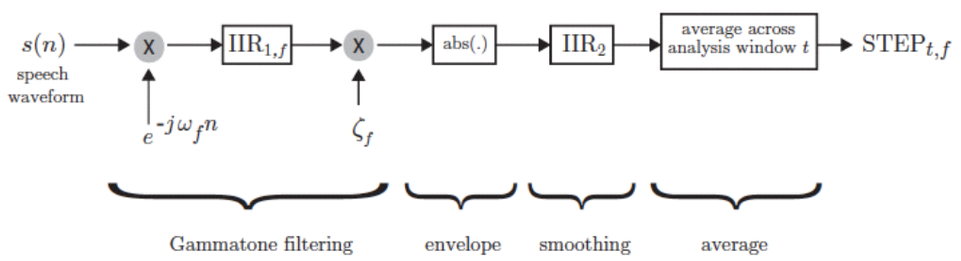
Baseline forced alignment

Proposed method

Utterance

Baseline forced alignment

STEP parametrisation

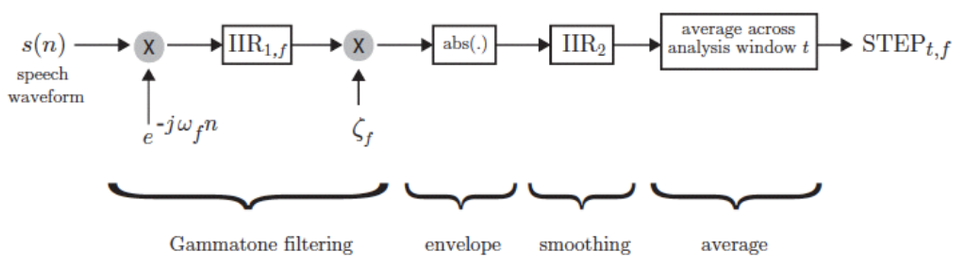


Proposed method

Utterance

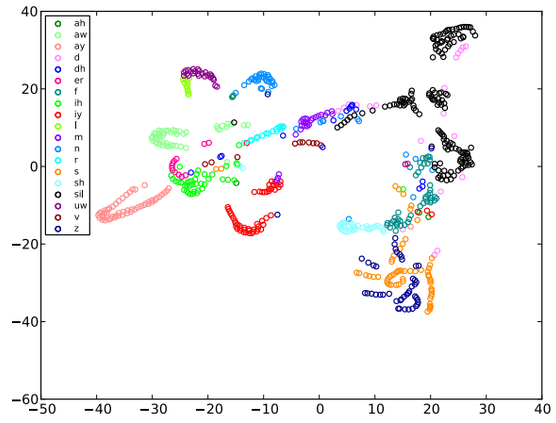
Baseline forced alignment

STEP parametrisation



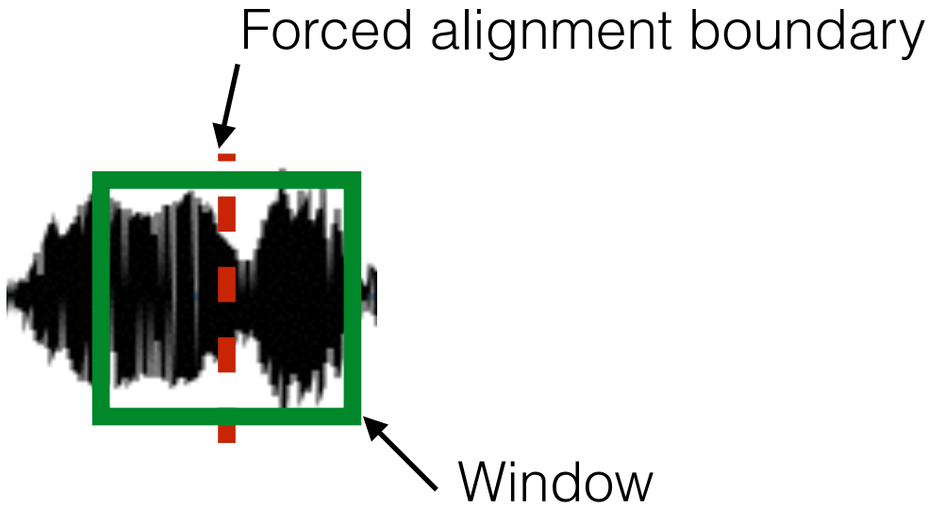
t-SNE on STEP

x5



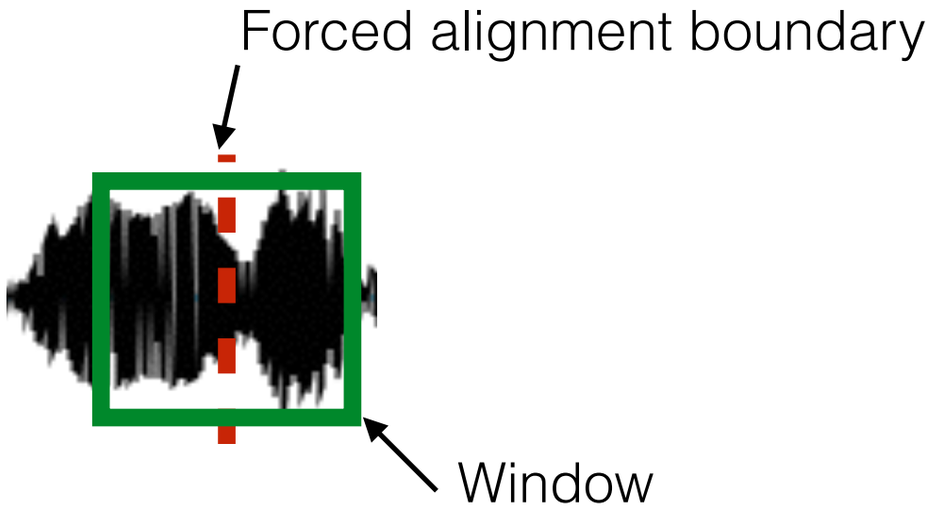
Proposed method

Window around baseline boundary

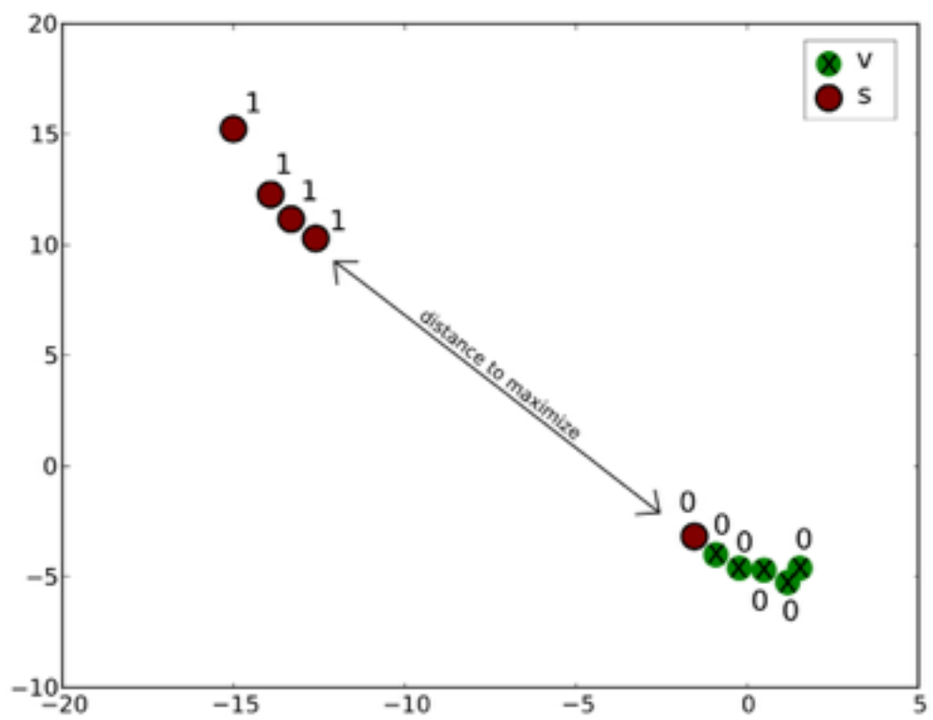


Proposed method

Window around baseline boundary

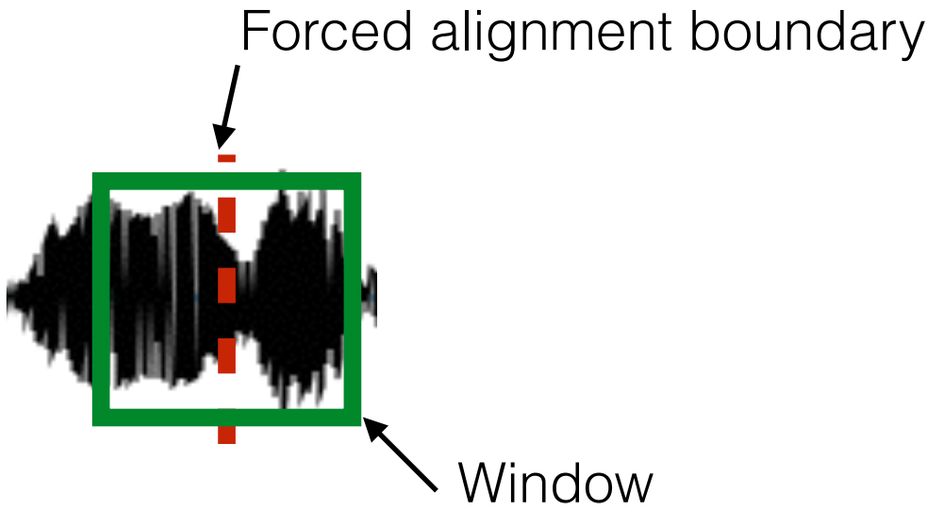


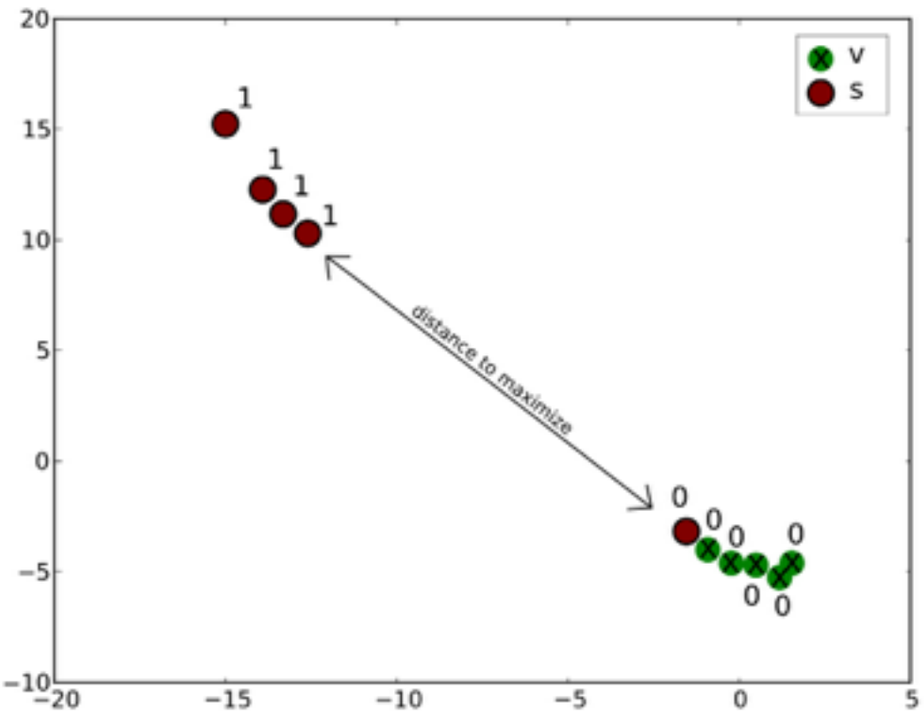
k-Means



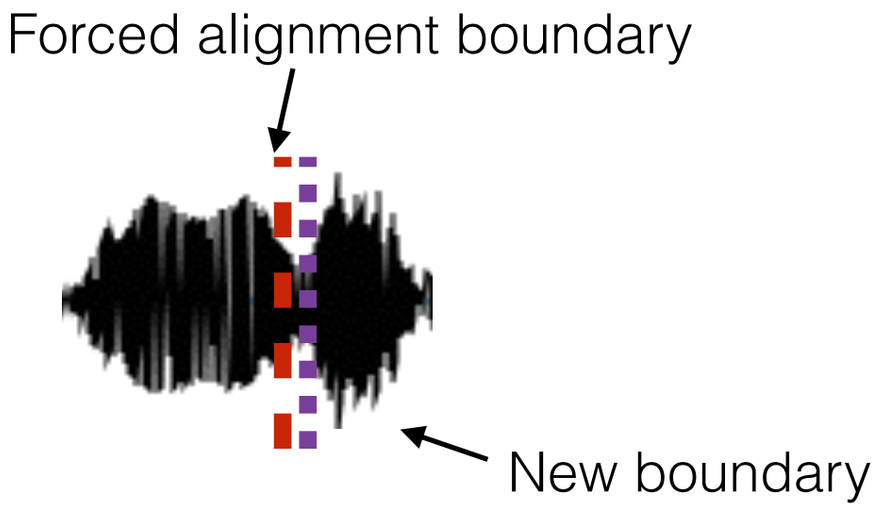
Proposed method

Window around baseline boundary





k-Means

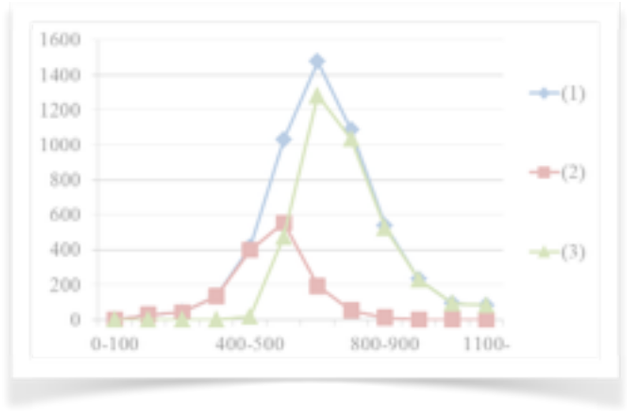


Evaluation and results

23.229%

M^r = 99.12

a/b = c/d



Evaluation

Dataset

- TIMIT dataset
- 5.5 hours of recordings of phonetically-balanced prompted speech
- 630 speakers in 8 major dialects of American English.
- 16 kHz with a 16 bit resolution.
- the 61 phones used in TIMIT were mapped to the CMU Pronouncing Dictionary, resulting a set of 40 phones.
- the silence segment boundaries were excluded from the evaluation

Evaluation

Baseline forced alignment systems

- Three separate acoustic models:
  - standard 13 MFCCs with energy, delta and delta-deltas;
  - 34 STEP with energy, delta and delta-deltas;
  - MFCC + STEP representation: 34 STEP coefficients extracted and 13 MFCCs, plus their delta and delta-deltas.
- All acoustic models used a 5 state, left-to-right, context-independent HMM for each phone.

Results

Table 1. Forced alignment results using different feature sets

	Accuracy [%]			
System	**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

Results

Table 2. Segmentation results for different alignment systems

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

Results

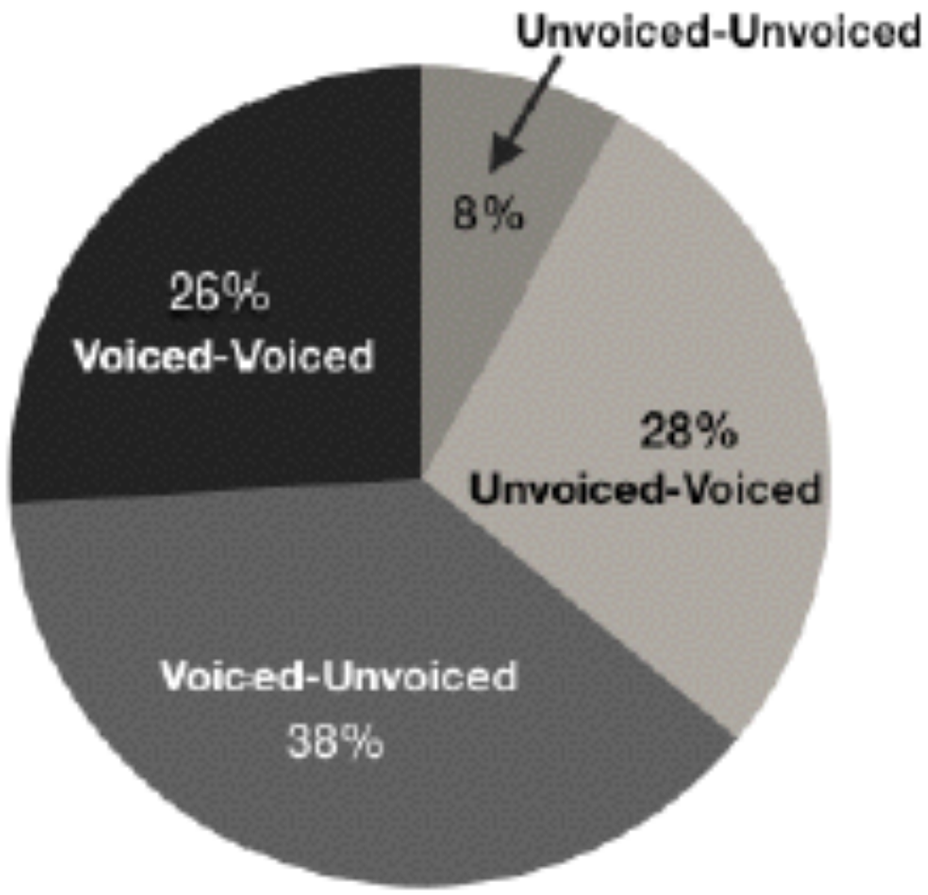


Fig. 4. TIMIT speech corpus division into phonetic categories

Results

Table 3. Results for **voiced** and **unvoiced** phonetic boundaries

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

Results

Table 4. Results for **unvoiced-voiced** and **voice-unvoiced** phonetic boundaries

Unvoiced-voiced phones

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-unvoiced phones

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

Results

Table 5. Results for **voiced-voiced** and **unvoiced-unvoiced** phonetic boundaries

Voiced-voiced boundaries

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

Unvoiced-unvoiced boundaries

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

Conclusions and future work

Conclusions

- the method can be applied to any speech resource in any language;
- better results for unvoiced phonemes, but worse for voiced phonemes;
- combination of t-SNE with some other feature reduction algorithm would be beneficial;
- ceiling effect in the case of unvoiced-unvoiced boundaries, where the baseline alignment even at a 50 ms threshold has an accuracy of only 60.86%;
- adjust representation and distance computation for each boundary type.

Thank you for your attention!

Adriana.Stan@com.utcluj.ro

http://speech.utcluj.ro/astan/

Q?

Acknowledgement

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] 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.
- [3] 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.
- [4] 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.
- [5] 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
- [6] S. Dusan and L. R. Rabiner, “On the relation between maximum spectral transition positions and phone boundaries.” in Proc. Interspeech, September 2006.
- [7] R. Prasad and B. Yegnanarayana, “Acoustic segmentation of speech using zero time liftering (ZTL),” in Proc. Interspeech, pp. 2292–2296, August 2013.
- [8] O. Kalinli, “Combination of auditory attention features with phone posteriors for better automatic phoneme segmentation.” In Proc. Interspeech, pp. 2302–2305, August 2013.