大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

INTERSPEECH 2016
San Francisco, California, USA

Enhance the word vector with prosodic information for the recurrent neural network based TTS system

Xin WANG, Shinji TAKAKI, Junichi YAMAGISHI
National Institute of Informatics, Japan
2016-09-11

contact: wangxin@nii.ac.jp
critical comments, suggestions, discussion are welcome

CONTENTS

- Introduction
- Previous work
- Method of this work
- Experiments and results
- Conclusion

INTRODUCTION

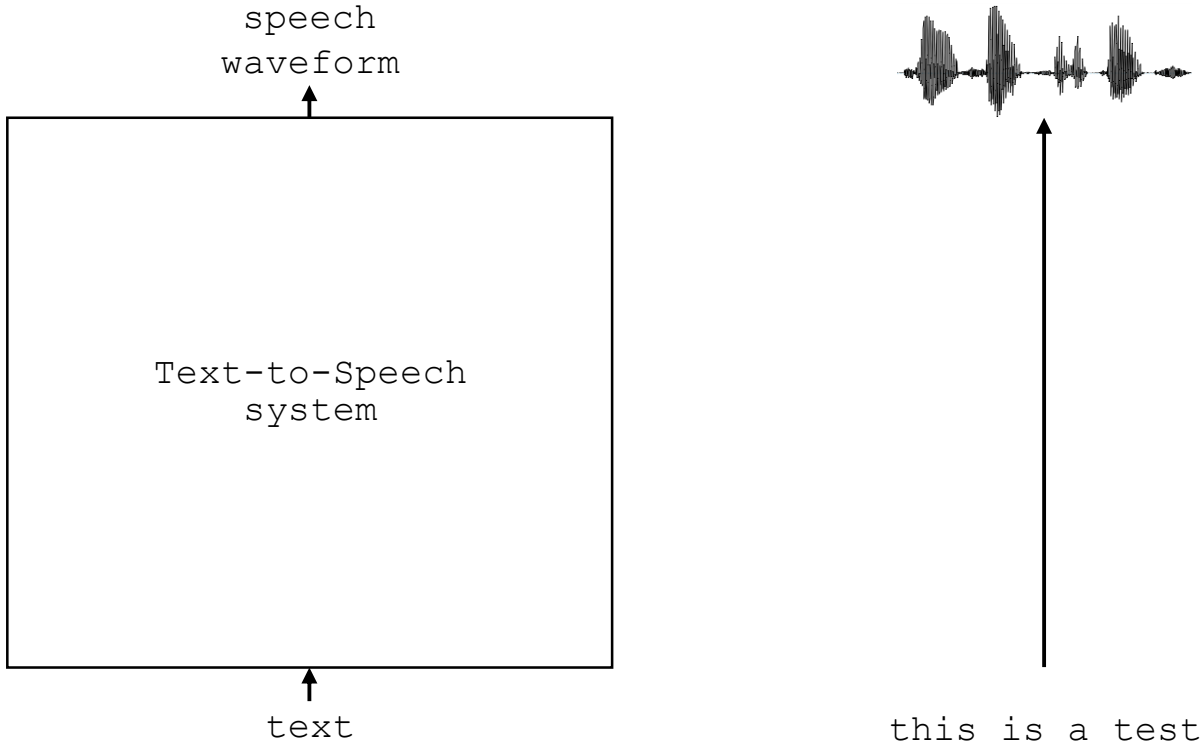
To avoid confusion

- "High Level Linguistic Features" ?
  - this work involves F0 trajectory, the surface string of word, and Tone and Break Indices (ToBI)
  - this work does NOT try to mine high level features (e.g. semantic)
- Prosody ?
  - this work use it to denote the super-segmental aspect of speech that is realized by the F0 trajectory

INTRODUCTION

Long-term goal

- End-to-End Text-to-Speech (TTS)



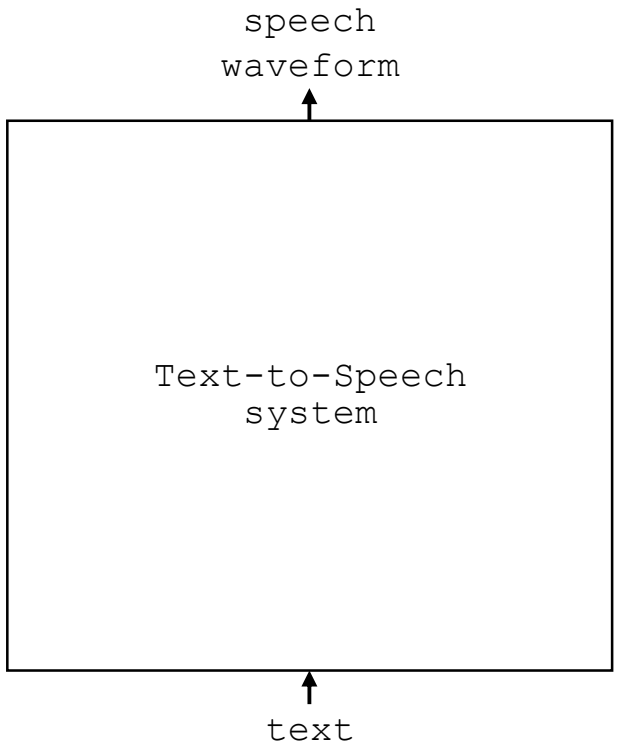
- *... creating a speech synthesizer for a new language or domain is too expensive, because current technology relies on **labelled data** and **human expertise**.* [1]

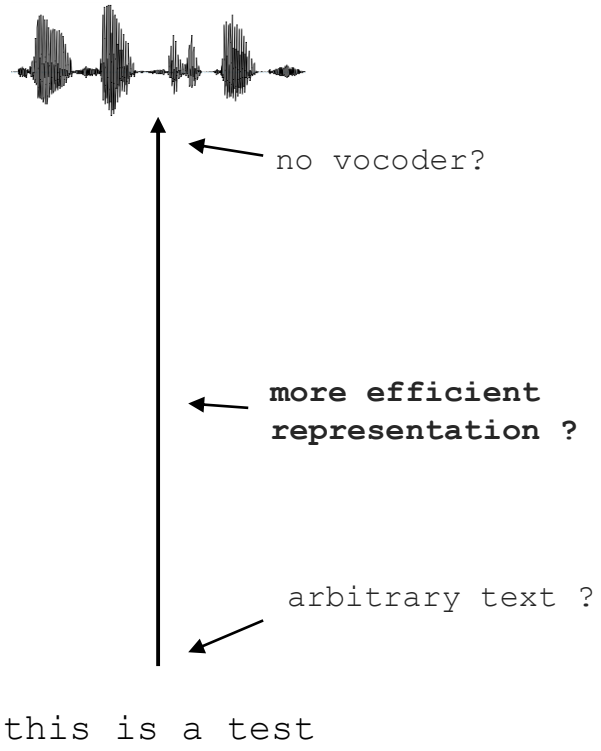
[1] Simple[4]All, http://simple4all.org

INTRODUCTION

Long-term goal

- End-to-End Text-to-Speech (TTS)



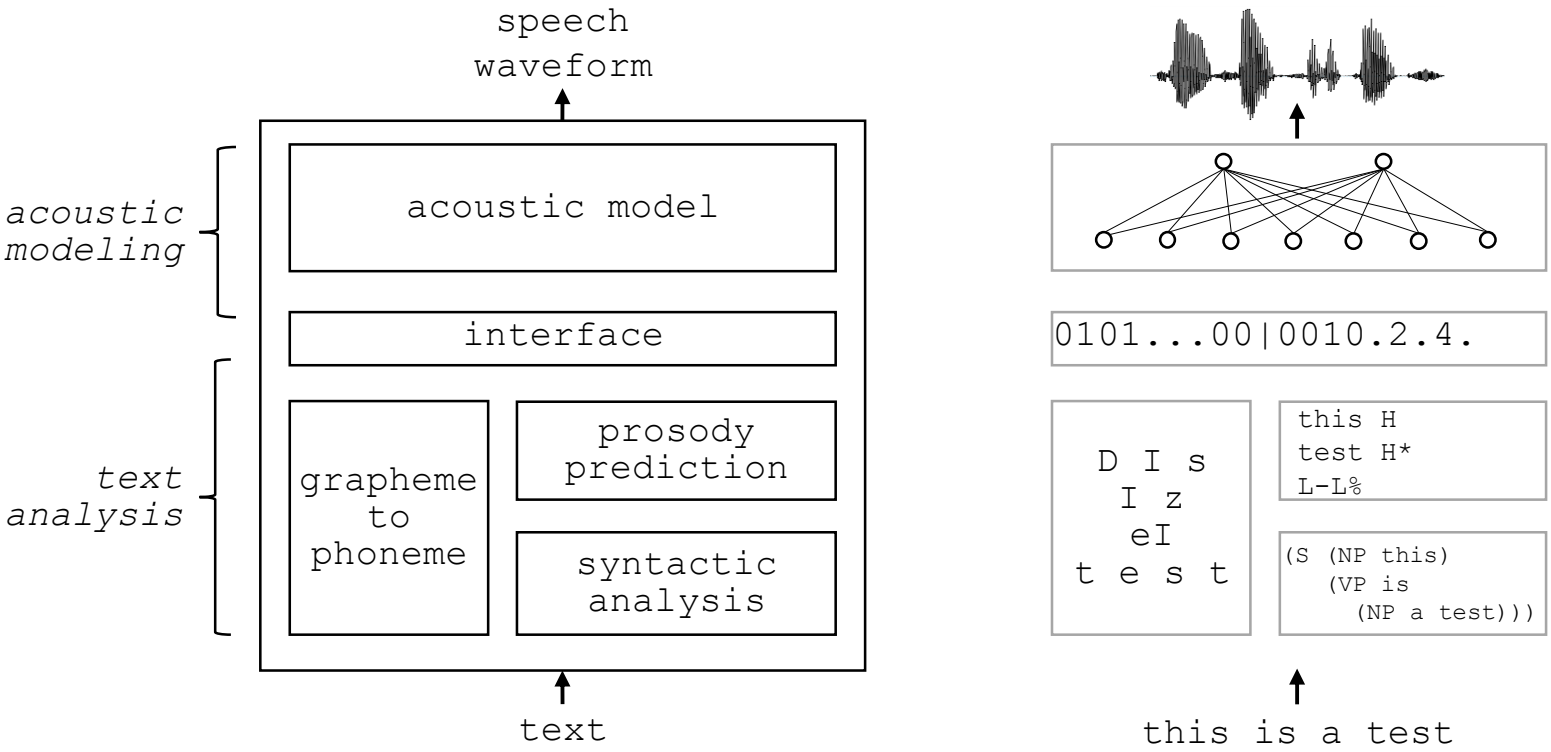


- efficient: less human expertise

INTRODUCTION

Starting point

- The common structure of TTS

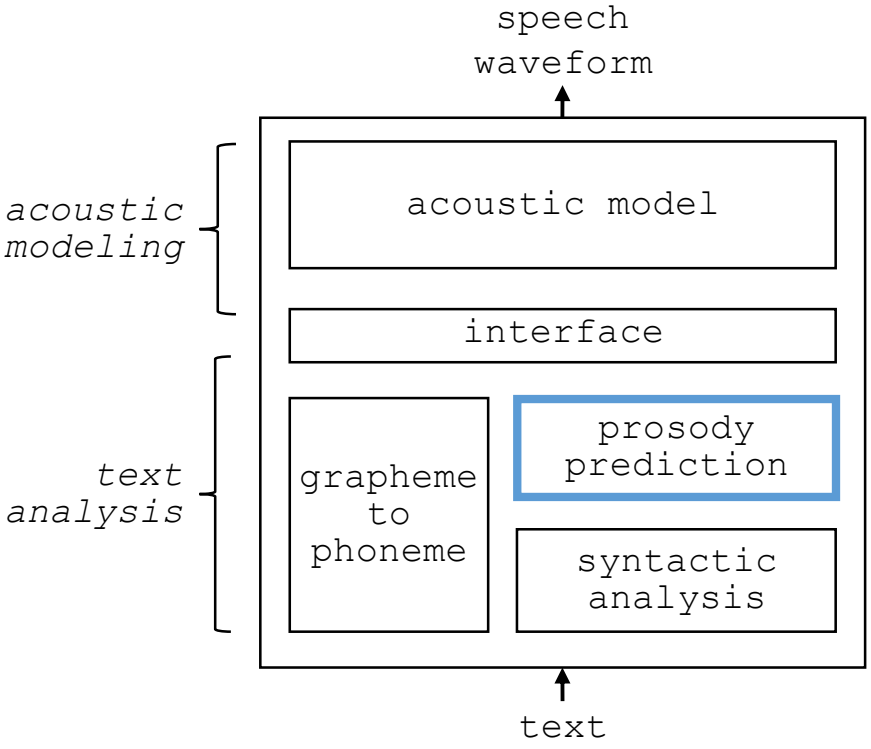


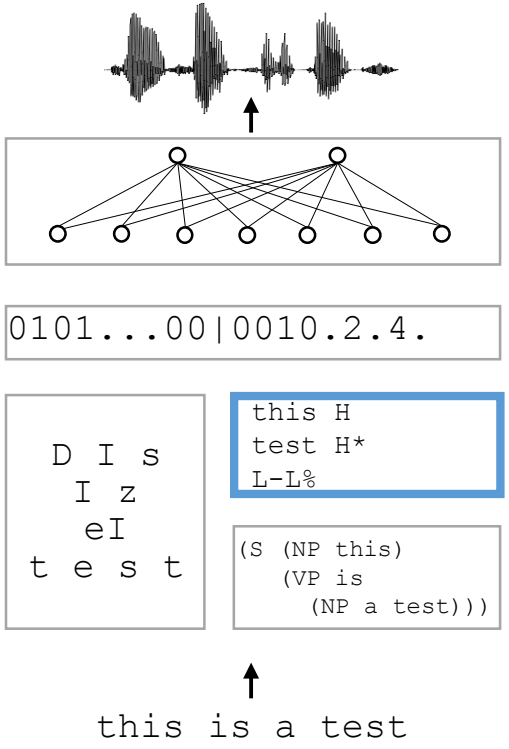
- front-end + back-end

INTRODUCTION

Question

- Efficient representation for prosody ?





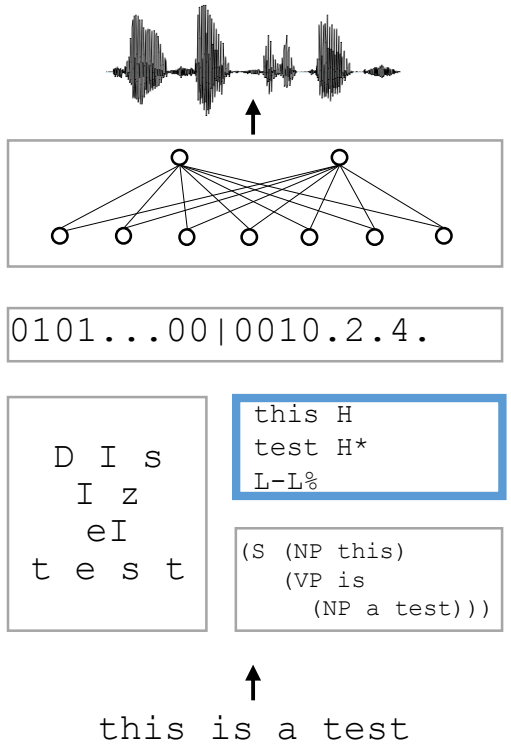
INTRODUCTION

Question

- Efficient representation for prosody ?

  - ToBI [2] is a common choice
  - ☹ a module to predict ToBI requires experts' annotation, which maybe 'expensive' on a larger corpus

  - Alternatives ?
  - ➢ unsupervised approach (previous work)
  - ➢ 'semi-supervised' approach (this work)



[2] Silverman, et. al. (1992). ToBI: a standard for labeling English prosody. In *ICSLP* (Vol. 2, pp. 867–870).
[3] Wightman, Colin W. (2002): "ToBI or not toBI?", In *Speech Prosody*, 25-29.

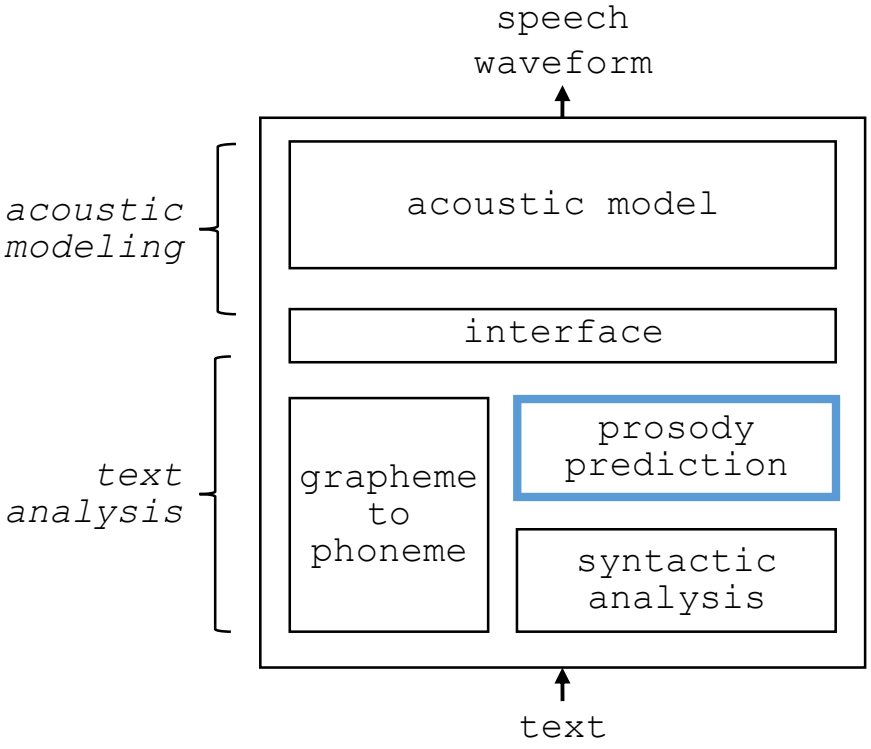
CONTENTS

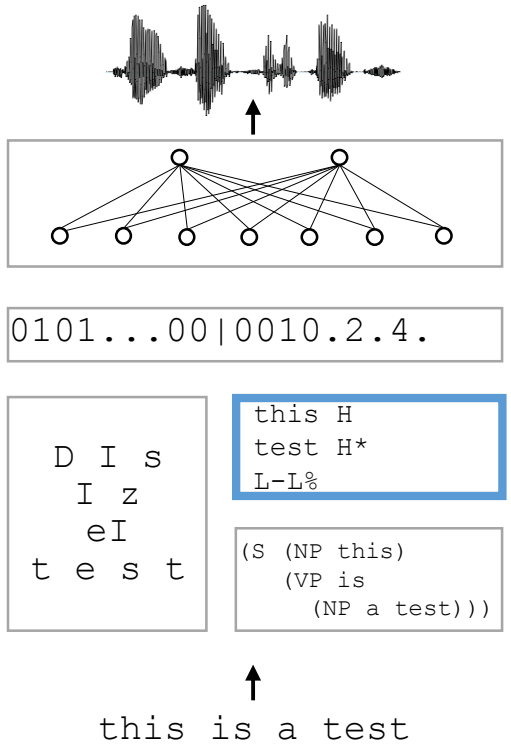
- Introduction
- Previous work
- Method of this work
- Experiments and results
- Conclusion

PREVIOUS WORK

TTS with word vectors

- Replace prosodic tags with word vectors

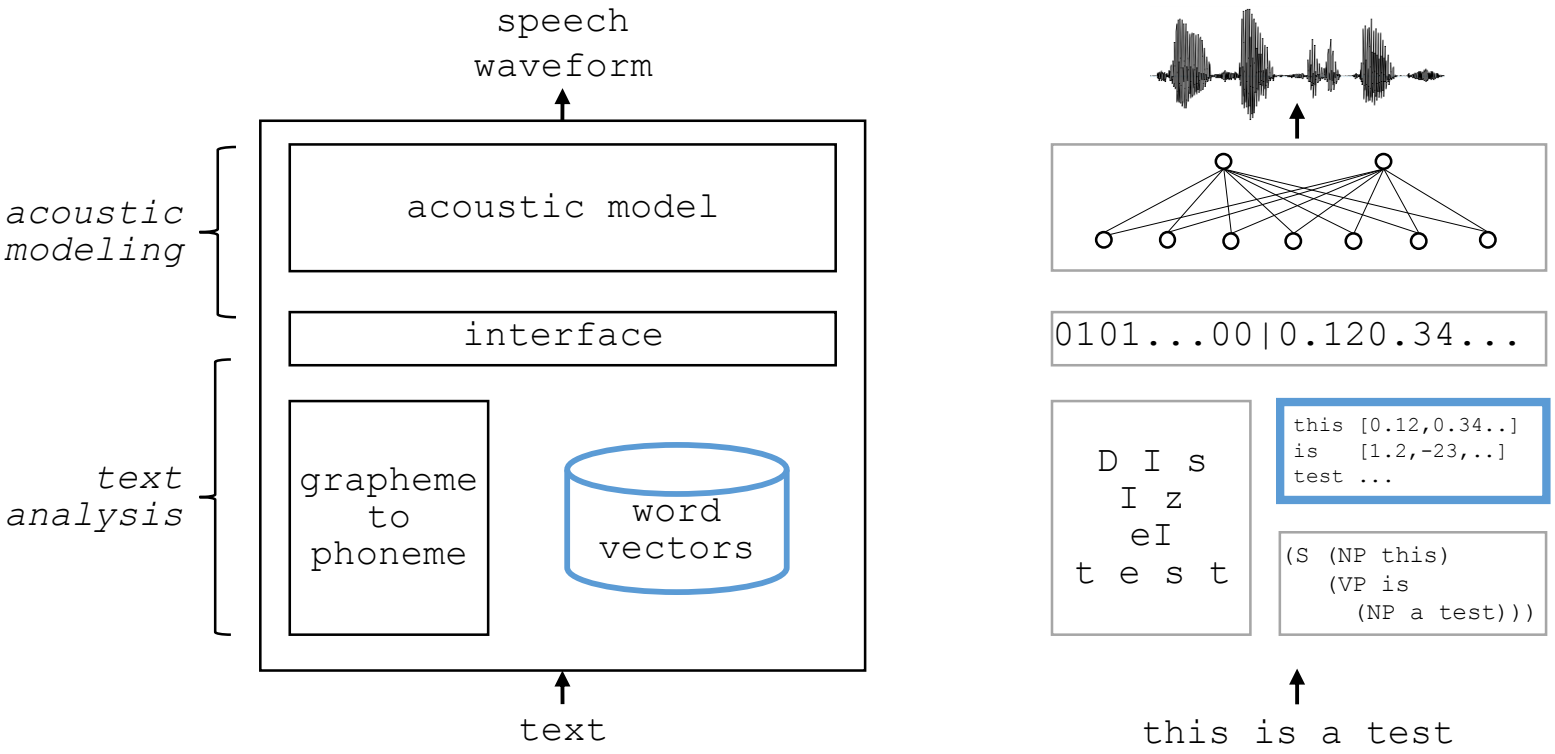




PREVIOUS WORK

TTS with word vectors

- Replace prosodic tags with word vectors [4]



- similar to the first work by another Wang [5]
- why word vectors [6] : **unsupervised** learning, linguistic regularity …

[4] Wang, X., Takaki, S., & Yamagishi, J. (2016). Investigation of Using Continuous Representation of Various Linguistic Units in Neural Network based TTS. *IEICE,* Vol.E99-D,No.10.
[5] Wang, P., Qian, Y., Soong, F. K., He, L., & Zhao, H. (2015). Word embedding for recurrent neural network based TTS synthesis. In *ICASSP* (pp. 4879-4883).
[6] Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In *HLT-NAACL* (pp. 746–751).

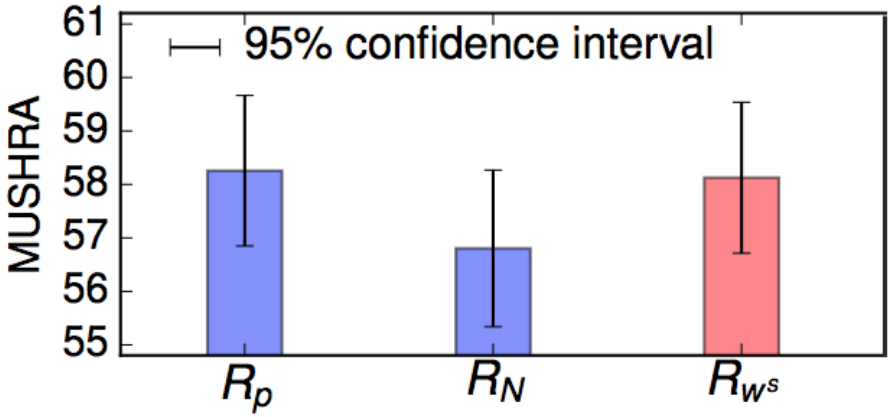
PREVIOUS WORK

TTS with word vectors

- Results of previous work [4]
  - Blizzard Challenge 2011, Nancy voice
  - Mushra test with 20 paird native speakers in CSTR

Tab1. Systems

ID	input to the acoustic model (a recurrent neural network)
R_p	phonemes + predicted prosodic tags
R_N	phonemes
R_{w^s}	phonemes + word vector



Further improvement ?

- unsupervised approach --> semi-supervised approach

  raw word vectors --> enhanced vectors with TTS-related information

[4] Wang, X., Takaki, S., & Yamagishi, J. (2016). Investigation of Using Continuous Representation of Various Linguistic Units in Neural Network based Text-to-Speech Synthesis. *IEICE,* Vol.E99-D,No.10.

CONTENTS

- Introduction
- Previous work
- Method of this work
- Experiments and results
- Conclusion

METHOD

Motivation

- Any reason to enhance the word vector for a specific task ?
  - it is based on unsupervised learning,
    and it only captures the topical similarity:

    "*coffee* and *cup* are more 'similar' than *car* and *train*" [7]

    and topical similarity may be insufficient for a specific task
    - e.g., predicting properties of concrete nouns [8]
      - ✓ taxonomic `a swan is an animal`
      - ✗ attributive `a swan is white`

[7] Hill, F. et.al. (2015). Simlex-999: Evaluating Semantic Models with Genuine Similarity Estimation. *Comput. Linguist.*, *41*(4), 665–695.
[8] Rubinstein, D. et.al. (2015). How well do distributional models capture different types of semantic knowledge? *ACL* (pp. 726–730).

METHOD

Solution

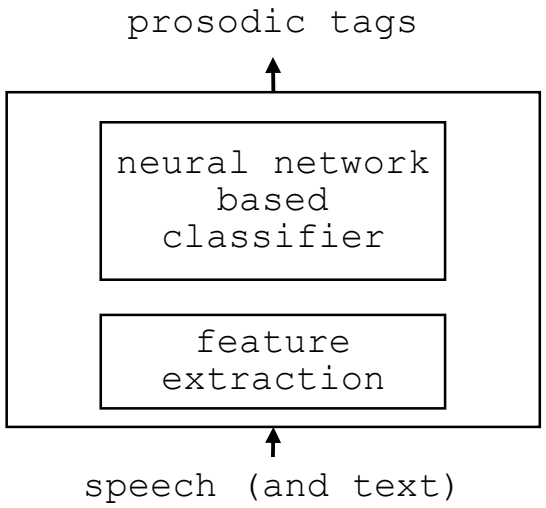
- Enhance word vectors using task-specific information
  - a semantic tagging task: semantic lexicon [9]
  - a syntactic parsing task: syntactic context [10]
- For TTS: enhance the vectors with prosodic information
  - where can we find the prosodic information ?
  - how can we enhance the vectors ?

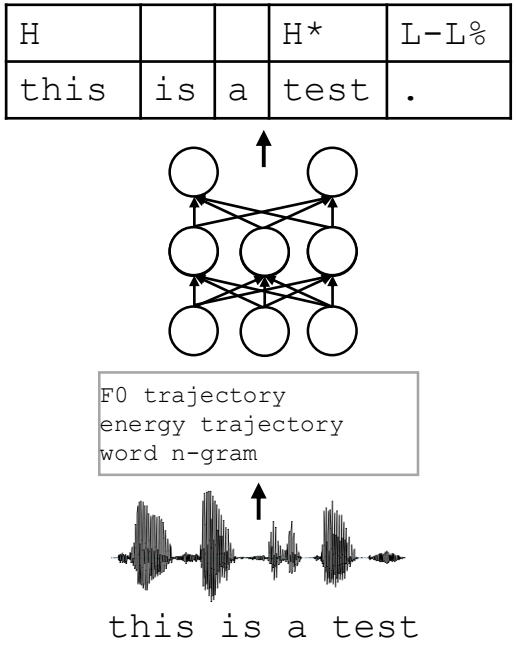
[9] Xu, C. et.al. (2014). RC-NET: A general framework for incorporating knowledge into word Representations. *CIKM* (pp. 1219–1228).
[10] Levy, O., & Goldberg, Y. (2014). Dependency-Based Word Embeddings. ACL (pp. 302–308).

METHOD

Enhance the word vector with prosodic information

- Where to find prosodic information ?
  - an auxiliary task: automatic prosodic annotation [11]



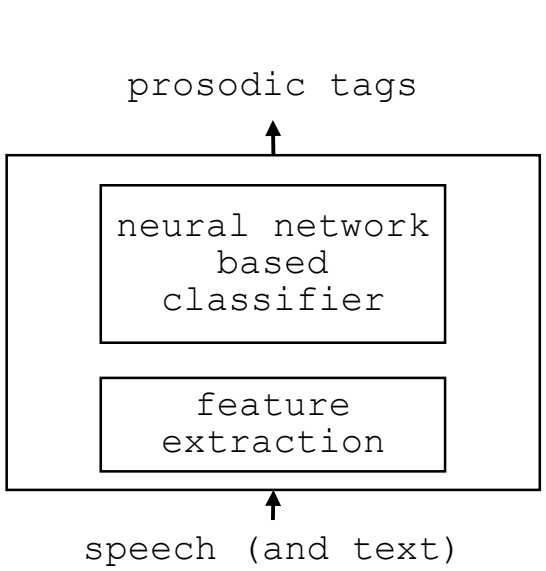


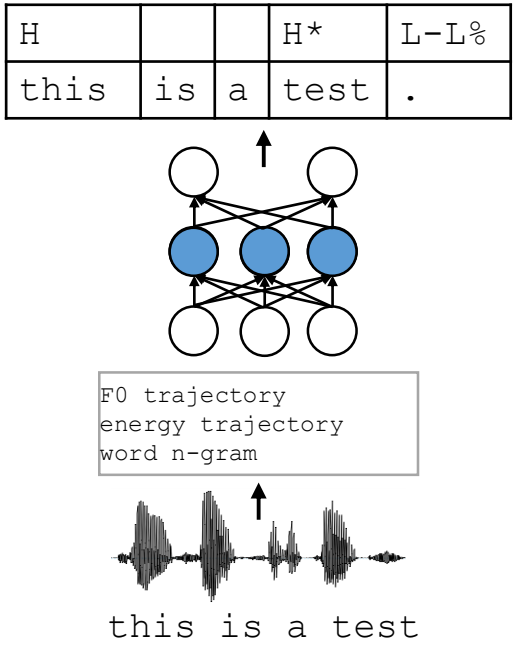
[11] Wightman, C. W., & Ostendorf, M. (1994). Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, *2*(4), 469–481.

METHOD

Enhance the word vector with prosodic information

- Where to find prosodic information ?
  - an auxiliary task: automatic prosodic annotation



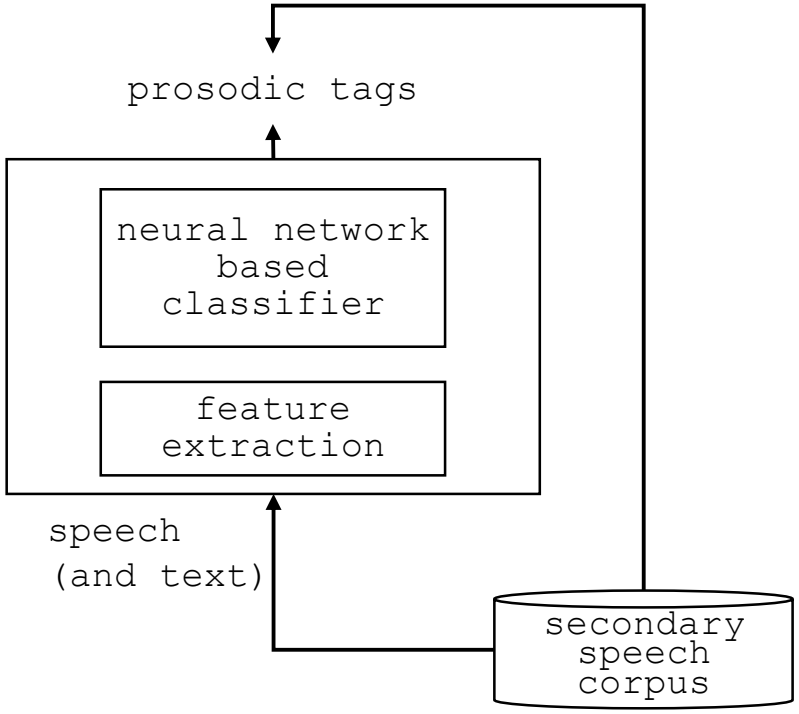


  - the output of the hidden layer of a neural network prosodic annotator

METHOD

Enhance the word vector with prosodic information

- Where to find prosodic information ?
  - an auxiliary task: automatic prosodic annotation



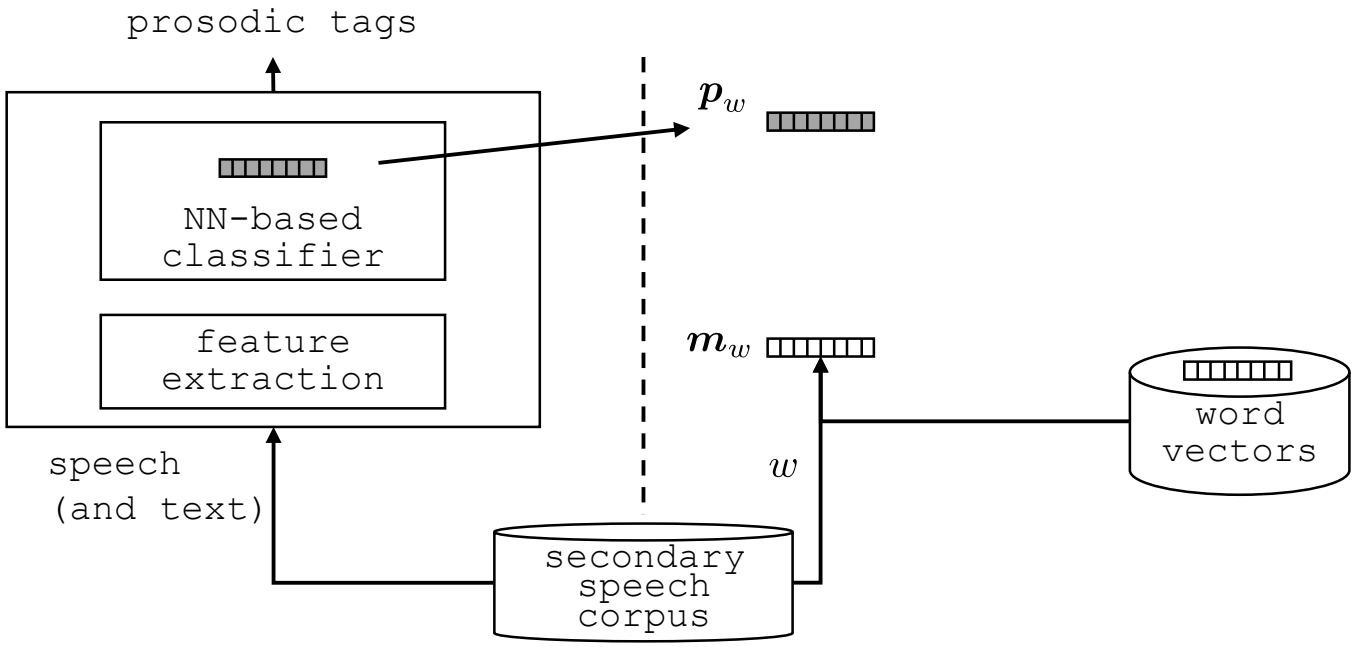
  - secondary corpus [12]: small, yet with expert prosodic annotation

[12] Watts, O. (2012). *Unsupervised Learning for Text-to-Speech Synthesis*. University of Edinburgh.

METHOD

Enhance the word vector with prosodic information

- How to enhance ?

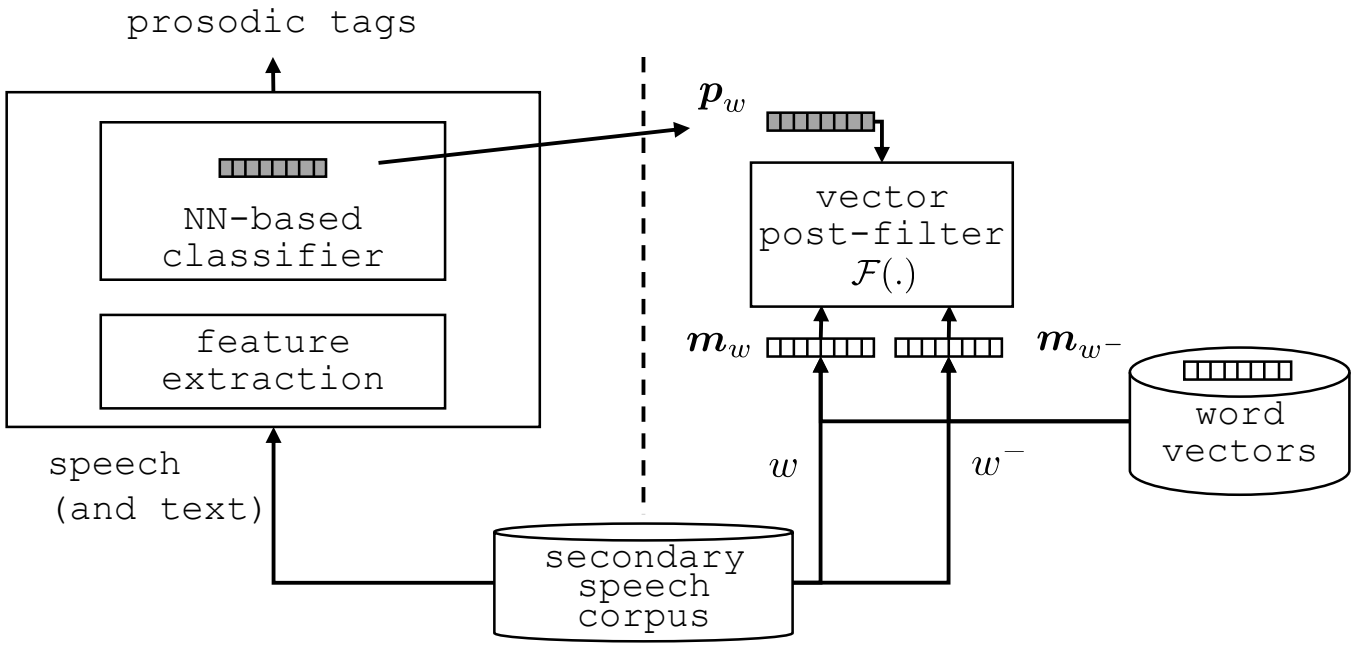


METHOD

Enhance the word vector with prosodic information

- How ? Train a post-filterer $\mathcal{F}(.)$ with triplet-ranking loss criterion [12]

$$E = \max\left[0, 1 - Sim(\boldsymbol{p}_w, \mathcal{F}(\boldsymbol{m}_w)) + Sim(\boldsymbol{p}_w, \mathcal{F}(\boldsymbol{m}_{w^-}))\right]. \qquad Sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||}$$

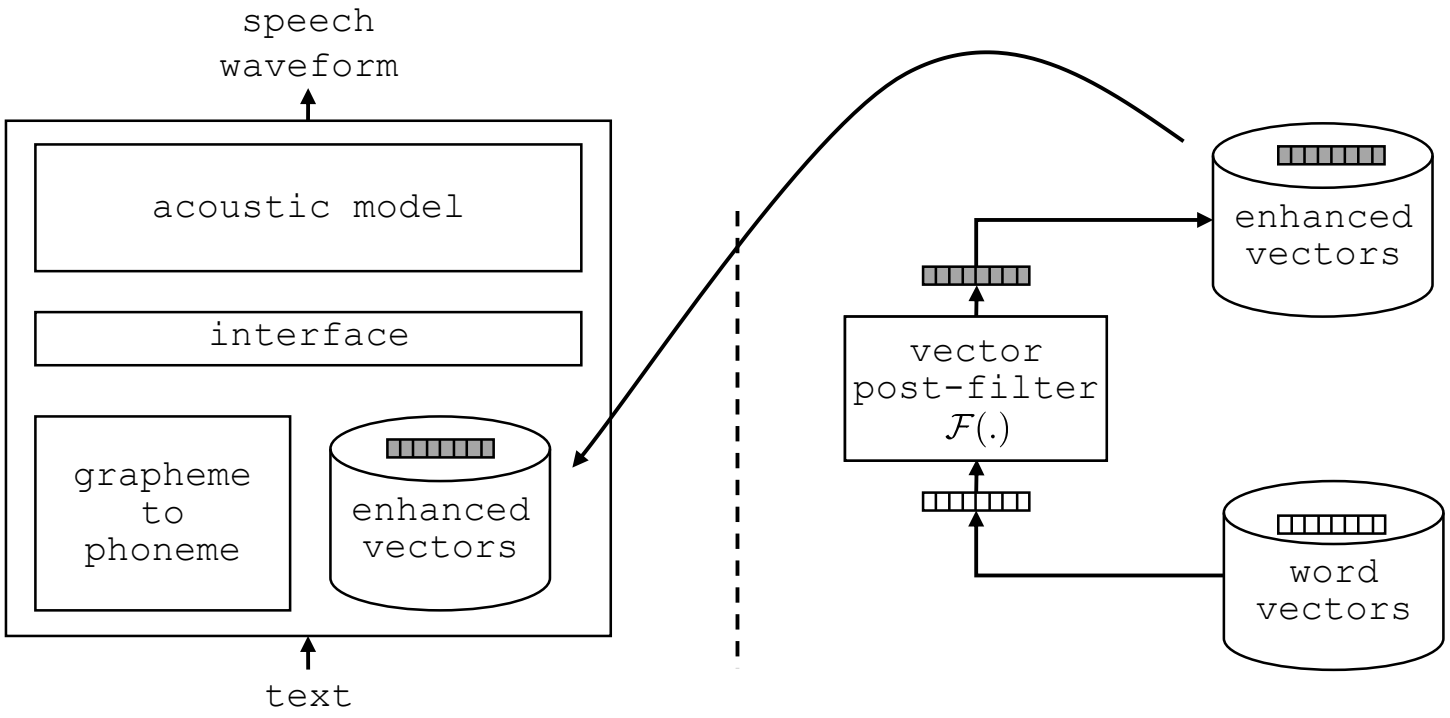


[12] Bengio, S., & Heigold, G. (2014). Word embeddings for speech recognition. In *INTERSPEECH-2014* (pp. 1053–1057).

METHOD

Enhance the word vector with prosodic information

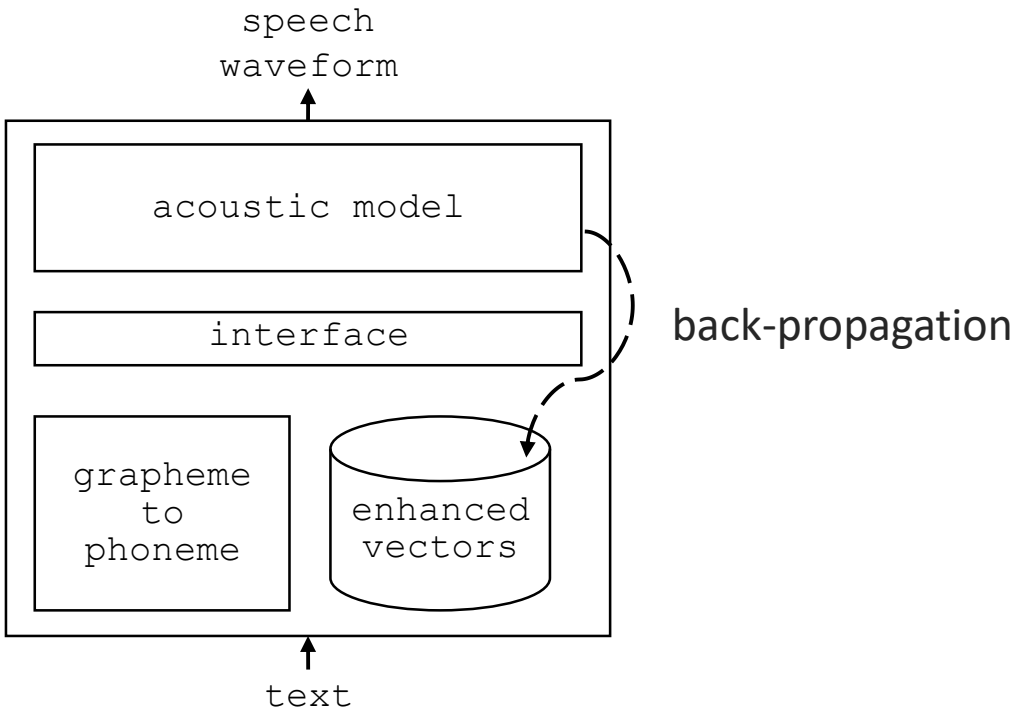
- Use enhanced vectors in TTS: a plug-in component



METHOD

Enhance the word vector with prosodic information

- Use enhanced vectors in TTS: fine-tune the vectors

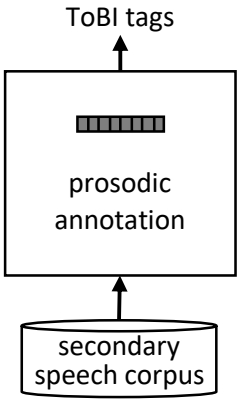


  - fine-tune to enhanced vectors after training the acoustic model

METHOD

Enhance the word vector with prosodic information

- Sum up



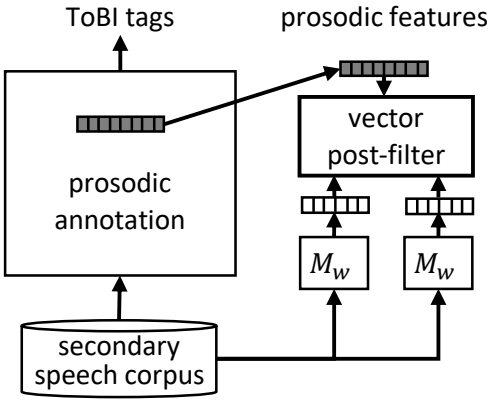
Post-filter training

- secondary corpus: small, with prosodic annotation

METHOD

Enhance the word vector with prosodic information

- Sum up



Post-filter training

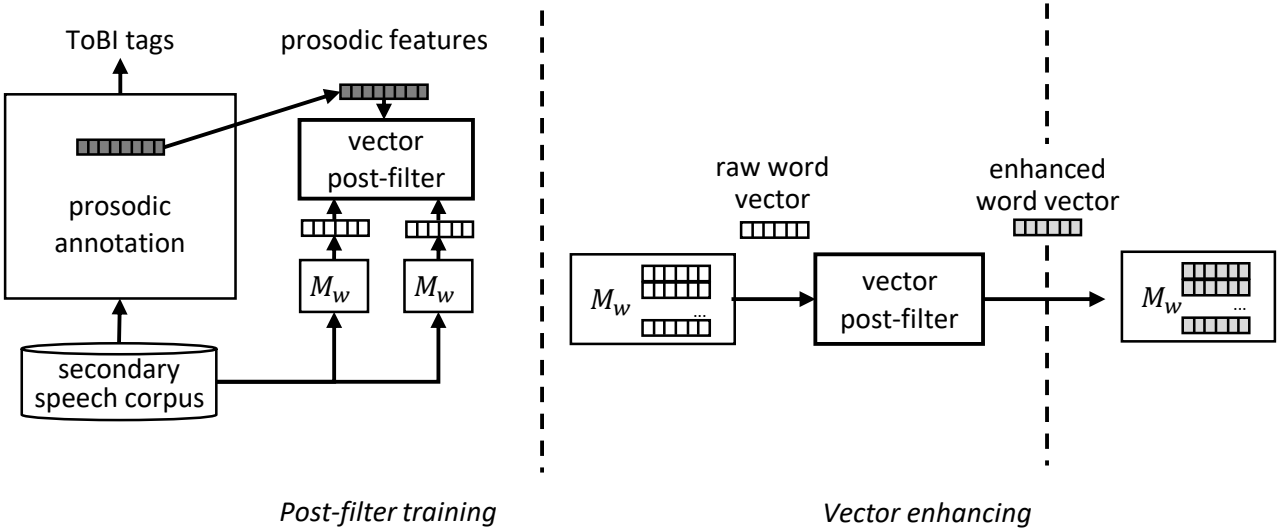
M is the set of word vector

- secondary corpus: small, with prosodic annotation

METHOD

Enhance the word vector with prosodic information

- Sum up

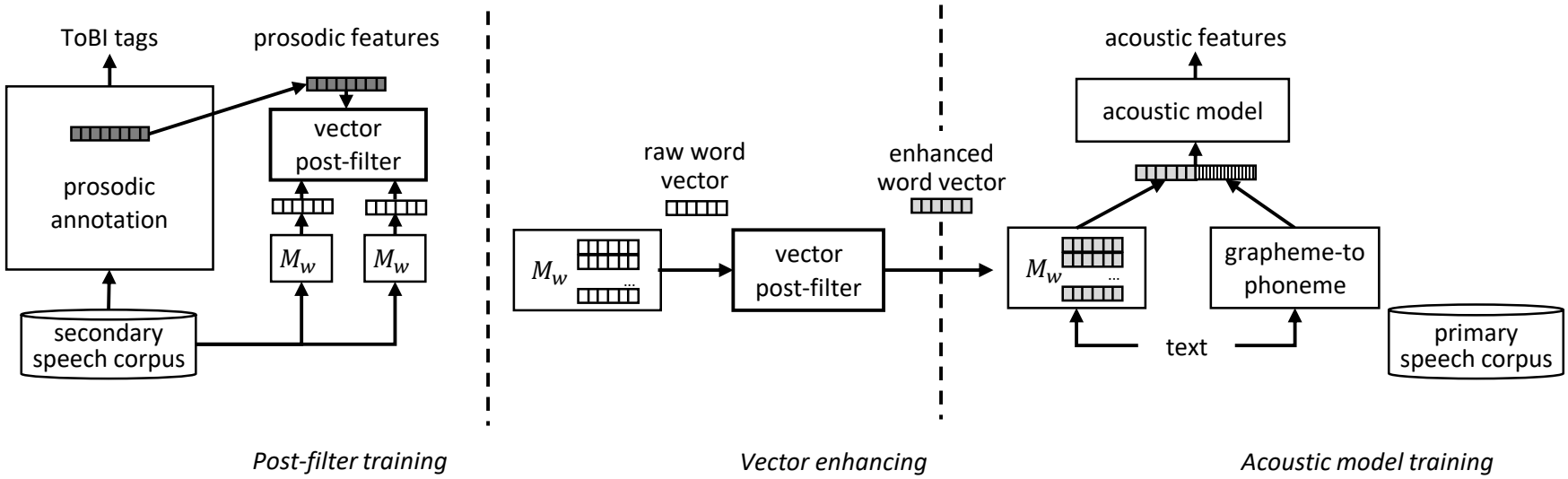


  - secondary corpus: small, with prosodic annotation

METHOD

Enhance the word vector with prosodic information

- Sum up

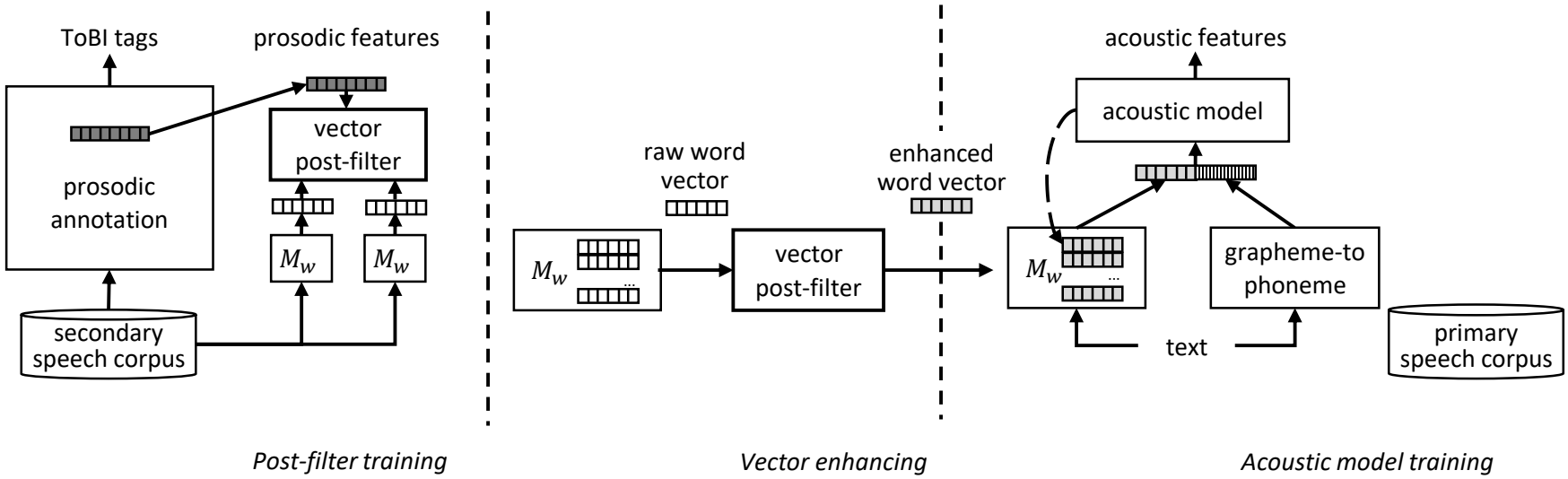


- secondary corpus: small, with expert prosodic annotation
- primary corpus: huge, w/o expert prosodic annotation

METHOD

Enhance the word vector with prosodic information

- Sum up



- secondary corpus: small, with expert prosodic annotation
- primary corpus: huge, w/o expert prosodic annotation

CONTENTS

- Introduction
- Previous work
- Method of this work
- Experiments and results
- Conclusion

EXPERIMENTS

Corpora and toolkit

- Primary corpus
  - 16 hours' recording of a female voice (BC2011, the Nancy voice [12])
- Secondary corpus
  - *f2b* set (< 1 hour) of Boston University Radio News Corpus [13]
- Raw word vectors
  - 80 dimensional vector (from RNN language model)
    http://www.fit.vutbr.cz/~imikolov/rnnlm/word_projections-80.txt.gz
- Toolkits
  - Acoustic model: modified CURRENNT [14] http://tonywangx.github.io
  - Prosodic annotation model: Theano [15]

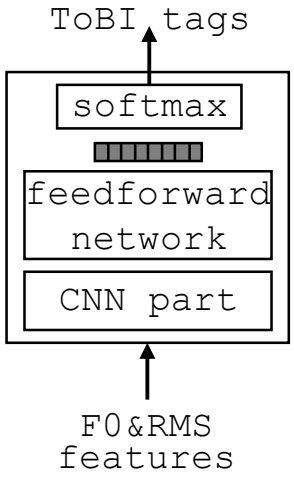
[12] King, S. & Karaiskos, V., 2011. The Blizzard Challenge 2011.
[13] Ostendorf, M., Price, P. J., & Shattuck-Hufnagel, S. (1995). The Boston University radio news corpus. *Linguistic Data Consortium*.
[14] Weninger, F., et. al. (2015). Introducing currennt: The munich open-source cuda recurrent neural network toolkit. *JMLR*, *16*(1), 547–551.
[15] Bastien, F., et. al. (2012). Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning Workshop

EXPERIMENTS

Automatic prosodic annotation model

- Input & target
  - input: wavelet representation of F0 (5 dim) [15] + RMS per frame (1 dim)
  - target: H*, !H*, L*, bi-tonal accent, other [16]

- Network structure
  - Convolutional neural network (CNN) part
    - 10 feature filters with size (5, 6)
    - max pooling stride (10, 1)
  - Feedforward network
    - 3 hidden layers with layer size 500 * 320 * 80



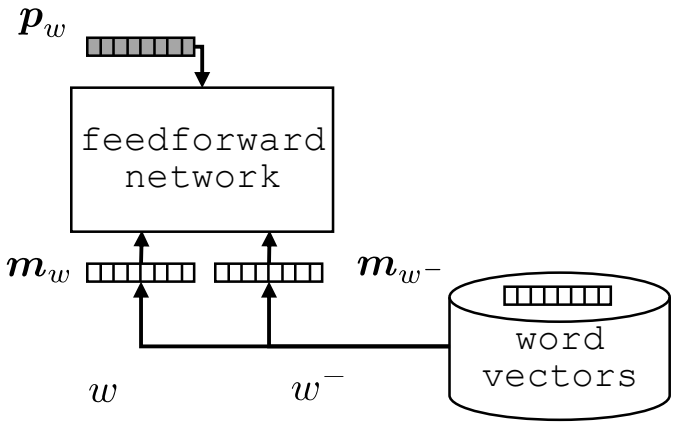
- Results on a binary annotation task (i.e. the word is accent or not)
  f_1 score: 0.882

[15] Suni, A. S., Aalto, D., Raitio, T., Alku, P., & Vainio, M. (2013). Wavelets for intonation modeling in HMM speech synthesis. SSW8, (pp 285-290).
[16] Black, A. W., & Hunt, A. J. (1996). Generating F0 contours from ToBI labels using linear regression. ICSLP (pp. 1385–1388)

EXPERIMENTS

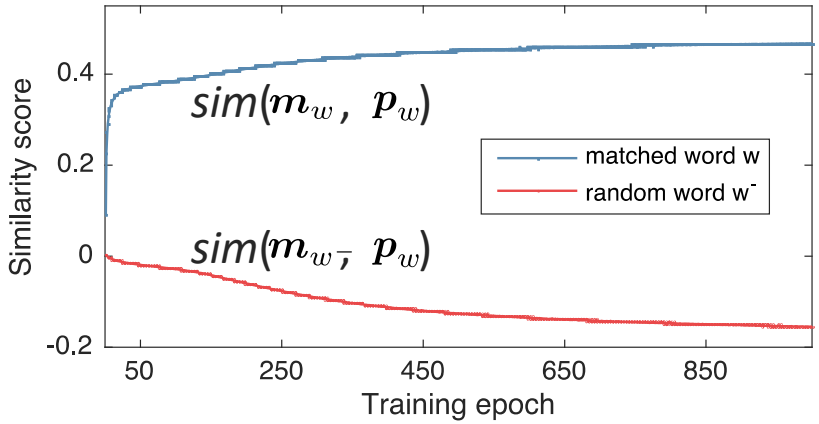
Vector poster-filter

- Model structure
  - 4 hidden layers with layer size: 80 * 160 * 160 * 80



- Results
  - similarity between target vector and input vector

$$Sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{y}||}$$



EXPERIMENTS

F0 generation in TTS

- Experimental systems

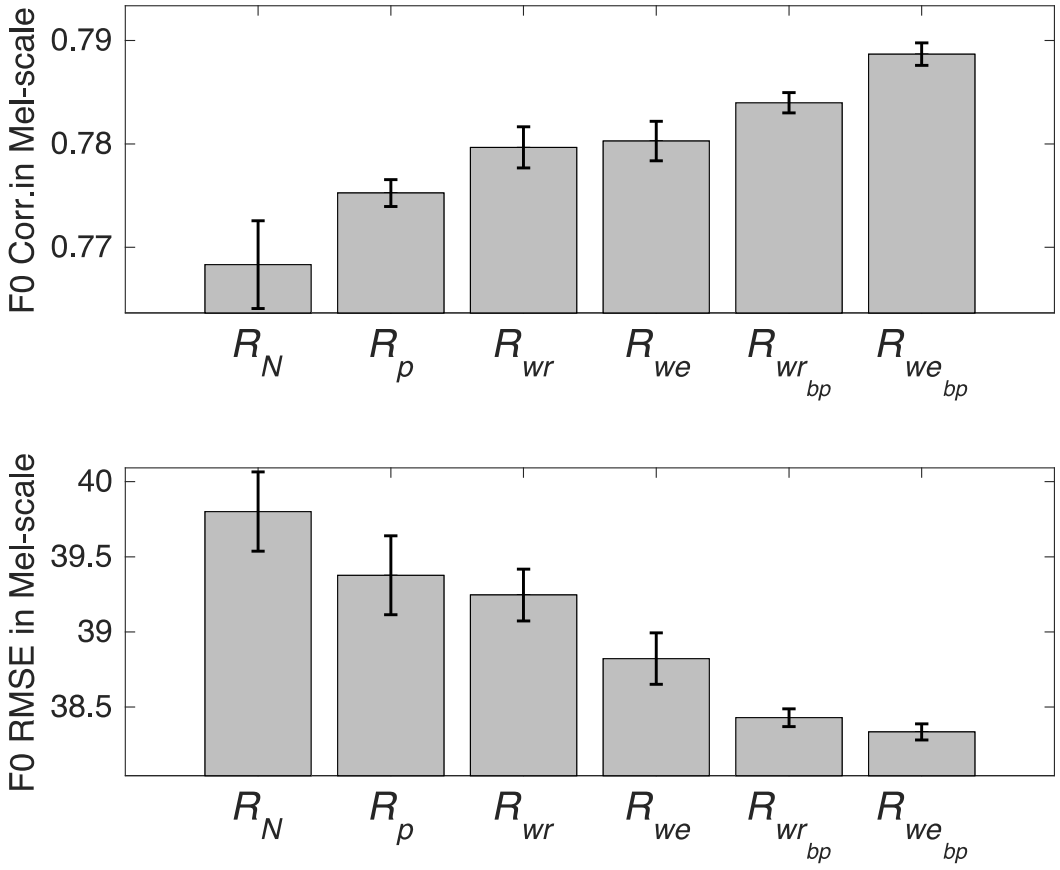
system ID	input to the acoustic model (F0 trajectory model)
R_N	phonemes
R_p	phonemes + conventional prosodic context (automatically predicted)
R_{wr}	phonemes + raw word vector
R_{we}	phonemes + enhanced word vector
$R_{wr_{bp}}$	phonemes + raw word vector tuned by back-propagation in TTS
$R_{we_{bp}}$	phonemes + enhanced word vector tuned by back-propagation in TTS

  - all systems use another acoustic model to predict spectral features

EXPERIMENTS

Results

- Objective test



R_N	only phoneme
R_p	+ prosodic context

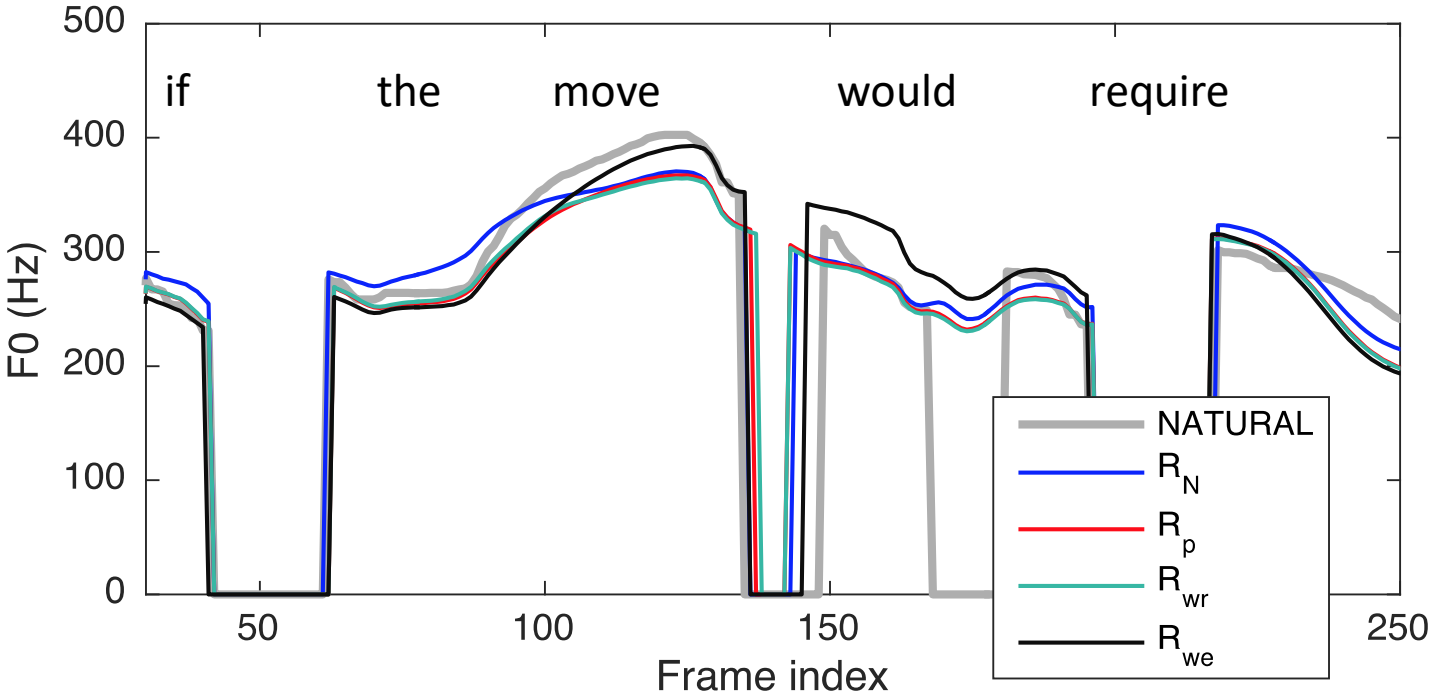
R_{wr}	+ raw word vector
R_{we}	+ enhanced word vector

$R_{wr_{bp}}$	+ raw word vector (fine tuned)
$R_{we_{bp}}$	+ enhanced word vector (fine tuned)

EXPERIMENTS

Results

- Sample



R_N	only phoneme
R_p	+ prosodic context

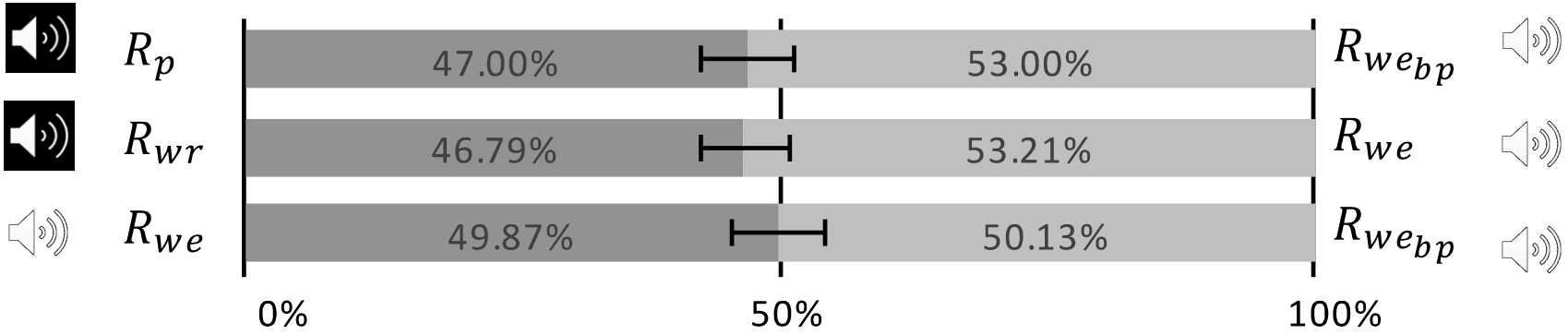
R_{wr}	+ raw word vector
R_{we}	+ enhanced word vecto

$R_{wr_{bp}}$	+ raw word vector (fine tuned)
$R_{we_{bp}}$	+ enhanced word vector (fine tuned

EXPERIMENTS

Results

- Subjective test
  - conducted in CSTR, by 20 paid native speakers



R_N	only phoneme
R_p	+ prosodic context

R_{wr}	+ raw word vector
R_{we}	+ enhanced word vector

$R_{wr_{bp}}$	+ raw word vector (fine tuned)
$R_{we_{bp}}$	+ enhanced word vector (fine tuned)

  - some evaluators favor R_{we} very much while others favor R_{wr} very much, because of missing the context of sentence ?
  - more samples: http://tonywangx.github.io
    https://www.dropbox.com/s/u1hxvqbpl5uj1r4/WE_0301_x01.tar.gz?dl=0

CONCLUSION

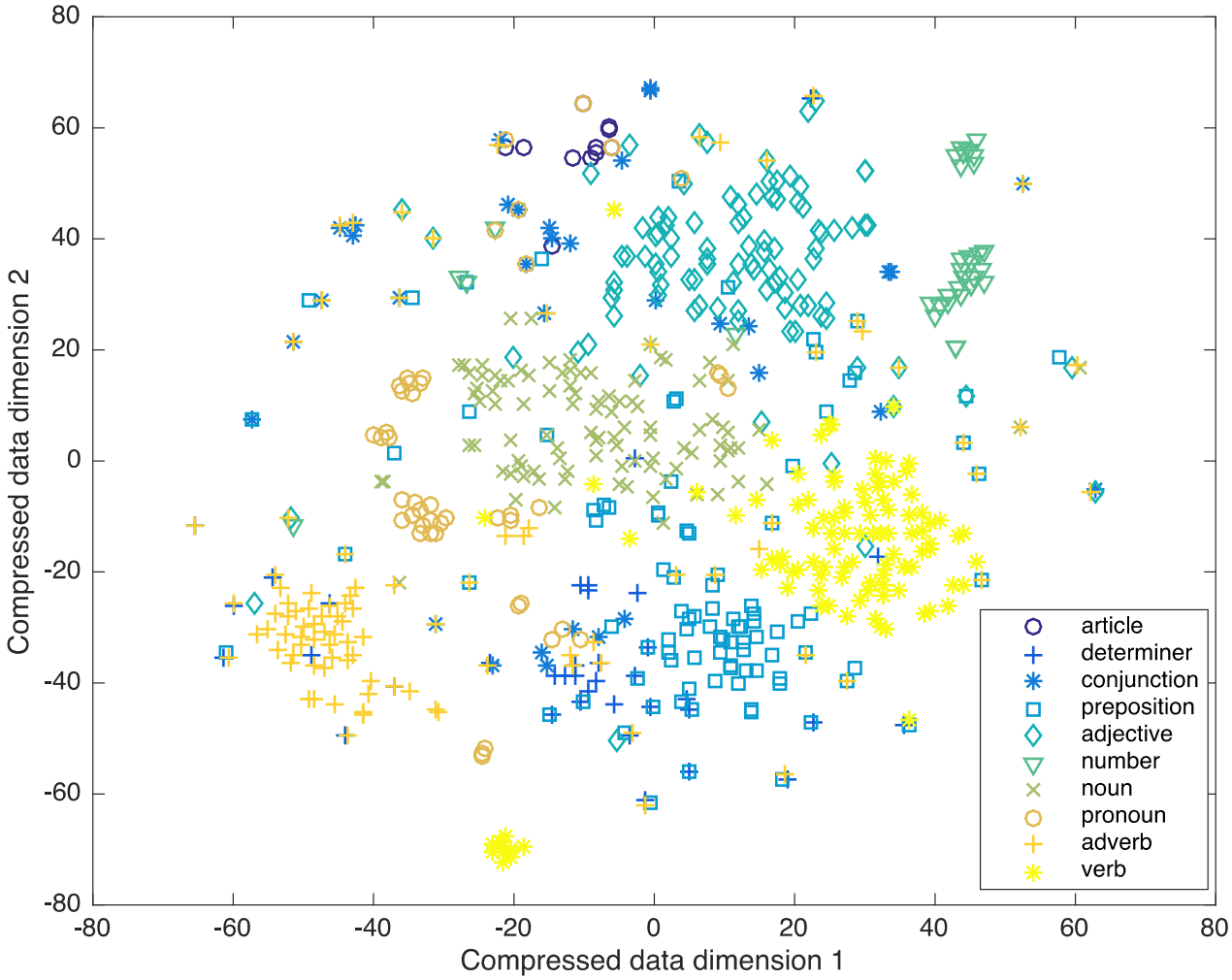
- Methods
  - enhance raw word vectors using prosodic features
    - prosodic features extracted from an prosodic annotation model
    - vector post-filter for enhancing raw vectors
- Results
  - improves the objective measure on F0 modelling
  - no significant improvement on perception
- Future work
  - alterative to modelling F0 trajectory at the frame level ?
  - annotation on the sub-set of primary corpus ?
  - use high level features (e.g., theme and rheme ?)

Thank you for your attention

Q & A

Visualizing the word vectors using *t*-SNE [16]

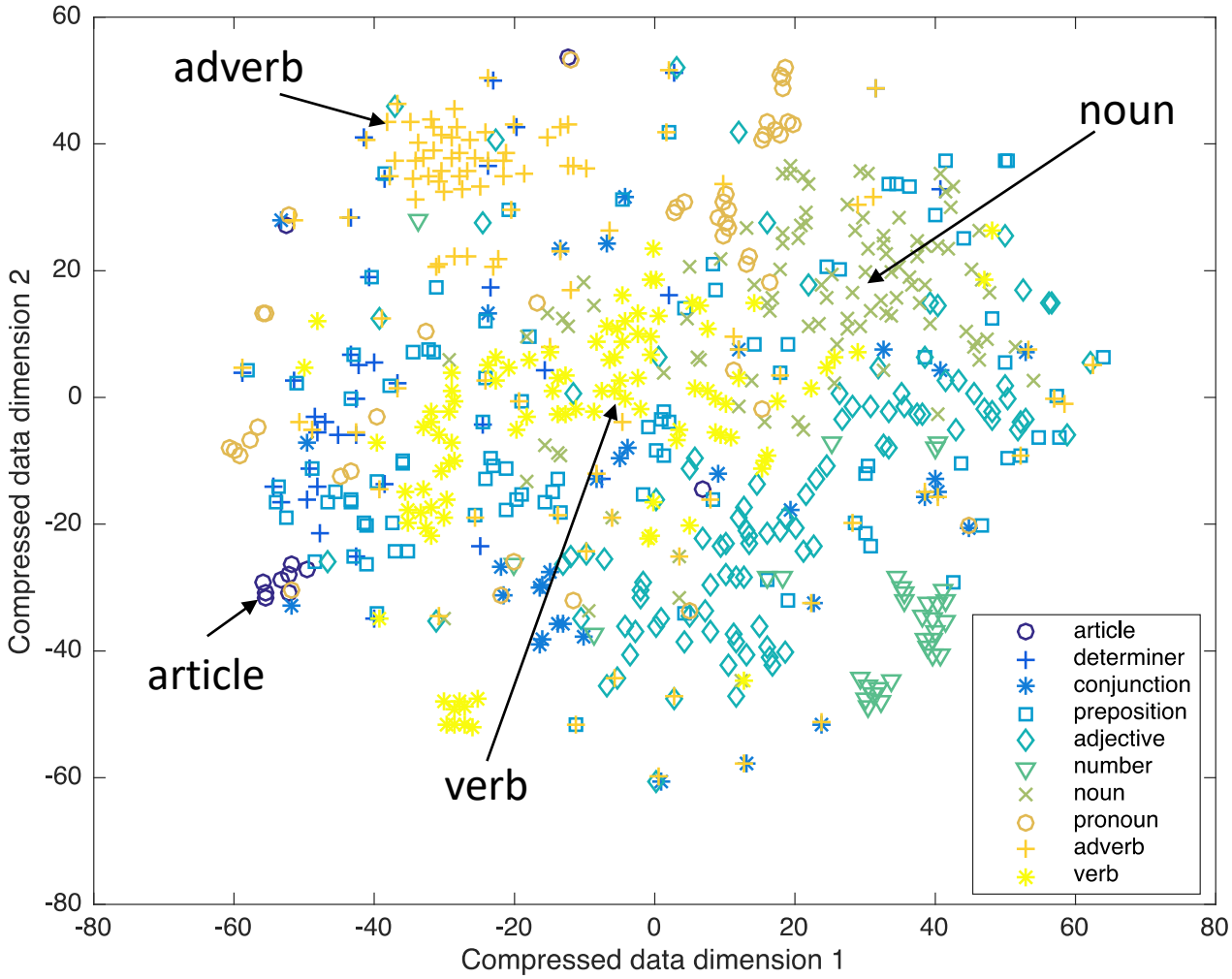
original word vectors http://www.fit.vutbr.cz/~imikolov/rnnlm/word_projections-80.txt.gz



[16] Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605. https://lvdmaaten.github.io/**tsne**/

Visualizing the word vectors using *t*-SNE [16]

word vectors after prosodic enhancement



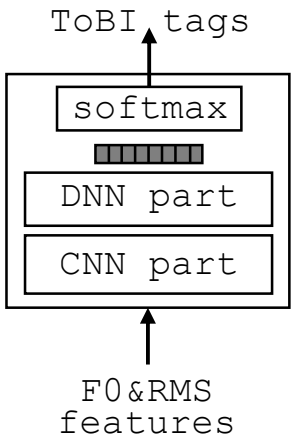
[16] Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605. https://lvdmaaten.github.io/**tsne**/

EXPERIMENTS

Prosodic annotation model

- Task
  - predict a ToBI tag for each word
  - input:
    - wavelet representation of F0 (5 dim)
    - RMS per frame (1 dim)
    - zero-padding to 160 frames/word
  - target: H*, !H*, L*, bitonal-accent, others

- Model structure
  - Convolutional neural network
    - 10 feature filters with size (5, 6)
    - max pooling stride (10, 1)
  - feedforward neural network
    - 500 * 320 * 80



PREVIOUS WORK

TTS with word vectors

- word vectors

	one-hot vector	word vector
cat	[0,0,0,…,0,1,0,…0]$\in \boldsymbol{R}^{\|V\|}$	[0.0023,…,0.0054…]$\in \boldsymbol{R}^{D}$
dog	[0,0,0,1,…,0,0,…0]$\in \boldsymbol{R}^{\|V\|}$	[0.0013,…,0.0033…]$\in \boldsymbol{R}^{D}$

- Advantages
  - linguistic regularity
  - low dimension (*D is* smaller than the size of dictionary $|V|$)
  - **unsupervised** learning based on plain text