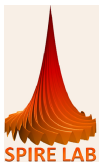
SPIRE-ABC: An online tool for acoustic-unit boundary correction (ABC) via crowdsourcing

Chiranjeevi Yarra, Kausthubha N K, and Prasanta Kumar Ghosh

SPIRE LAB
Electrical Engineering,
Indian Institute of Science (IISc), Bangalore, India



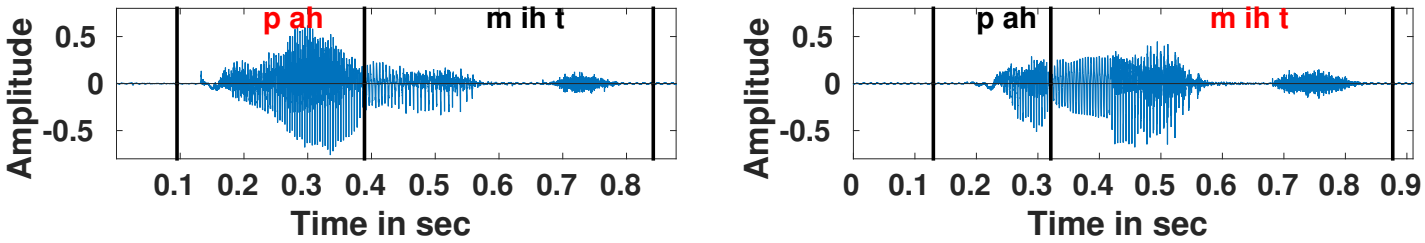
Introduction

- Need of time-aligned acoustic-unit (AU – Word, Syllable and Phoneme) boundaries[1]
  - Human computer interaction
  - Computer assisted language learning (CALL)

[1] Hönig, Batliner, and Nöth, "Automatic assessment of non-native prosody annotation, modelling and evaluation", 2012

[2] Franco et al., "Automatic detection of phone-level mispronunciation for language learning.", 1999

Introduction

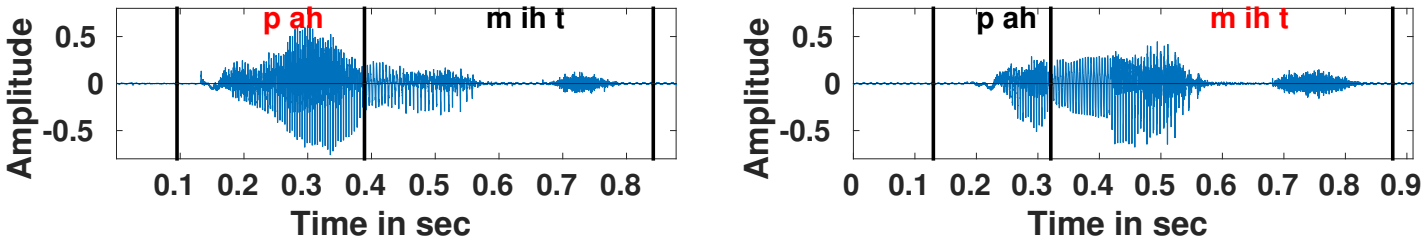


- Need of time-aligned acoustic-unit (AU – Word, Syllable and Phoneme) boundaries[1]
  - Human computer interaction
  - Computer assisted language learning (CALL)

[1] Hönig, Batliner, and Nöth, "Automatic assessment of non-native prosody annotation, modelling and evaluation", 2012

[2] Franco et al., "Automatic detection of phone-level mispronunciation for language learning.", 1999

Introduction

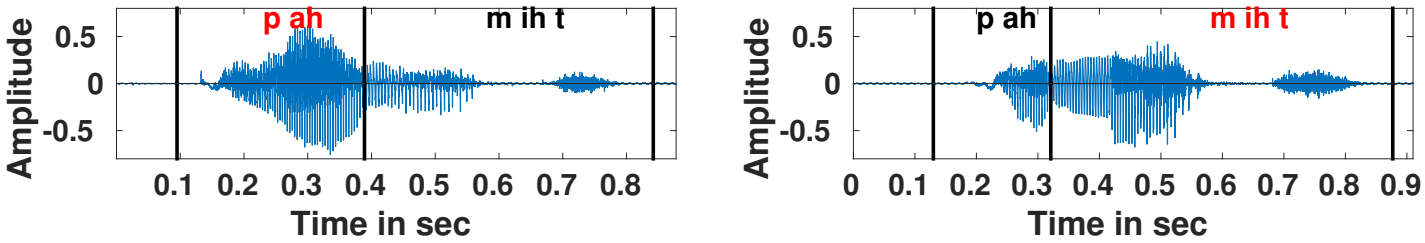


- Need of time-aligned acoustic-unit (AU – Word, Syllable and Phoneme) boundaries[1]
  - Human computer interaction
  - Computer assisted language learning (CALL)
- Typically, these boundaries are estimated using automatic speech recognition (ASR) system[2].

[1] Hönig, Batliner, and Nöth, "Automatic assessment of non-native prosody annotation, modelling and evaluation", 2012

[2] Franco et al., "Automatic detection of phone-level mispronunciation for language learning.", 1999

Introduction

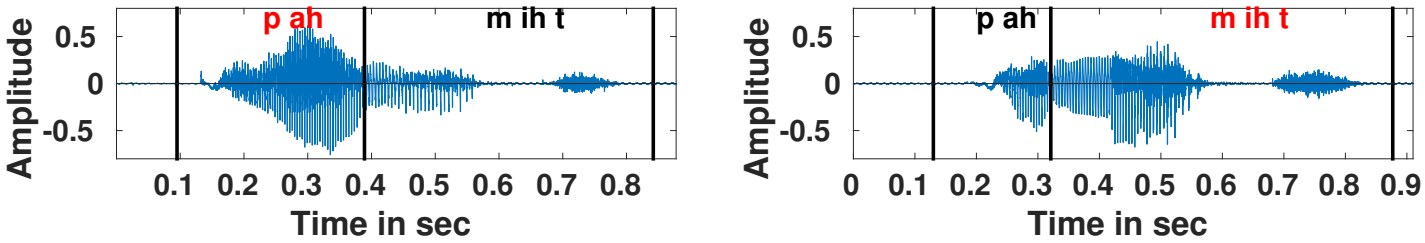


- Need of time-aligned acoustic-unit (AU – Word, Syllable and Phoneme) boundaries[1]
  - Human computer interaction
  - Computer assisted language learning (CALL)
- Typically, these boundaries are estimated using automatic speech recognition (ASR) system[2].
- However, these boundaries often suffer from errors due to inaccuracies in ASR system.

[1] Hönig, Batliner, and Nöth, "Automatic assessment of non-native prosody annotation, modelling and evaluation", 2012

[2] Franco et al., "Automatic detection of phone-level mispronunciation for language learning.", 1999

Introduction



- Need of time-aligned acoustic-unit (AU – Word, Syllable and Phoneme) boundaries[1]
  - Human computer interaction
  - Computer assisted language learning (CALL)
- Typically, these boundaries are estimated using automatic speech recognition (ASR) system[2].
- However, these boundaries often suffer from errors due to inaccuracies in ASR system.

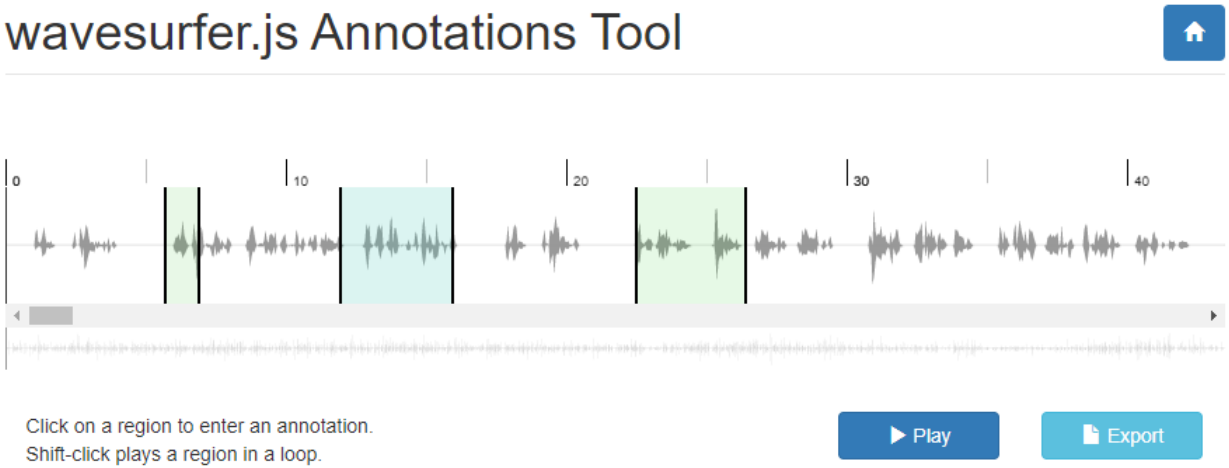
Goal of SPIRE-ABC

Facilitates the manual correction of AU boundaries (online) with naive annotators for cost-effective solutions

[1] Hönig, Batliner, and Nöth, "Automatic assessment of non-native prosody annotation, modelling and evaluation", 2012

[2] Franco et al., "Automatic detection of phone-level mispronunciation for language learning.", 1999

Existing online annotation tool[6]



- WaveSurfer is a general purpose JavaScript.

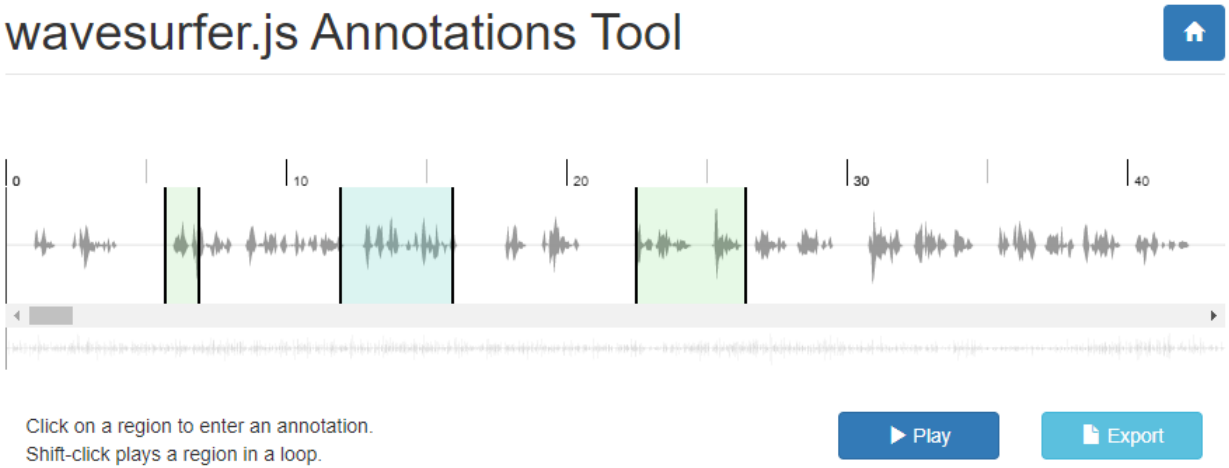
[3] Saiz, Matuszewski, and Goldszmidt, "Audio oriented UI components for the web platform", 2015

[4] Baker et al., "BioAcoustica: a free and open repository and analysis platform for bioacoustics", 2015

[5] Matuszewski, Schnell, and Goldszmidt, "Interactive Audiovisual Rendering of Recorded Audio and Related Data with the WavesJS Building Blocks", 2016

[6] Katspaugh, "wavesurfer.js", 2012

Existing online annotation tool[6]



- WaveSurfer is a general purpose JavaScript.
- It has been used via crowdsourcing in many applications include – 1) combining audios[3], 2) voice activity detection[4], and 3) audio rendering[5].

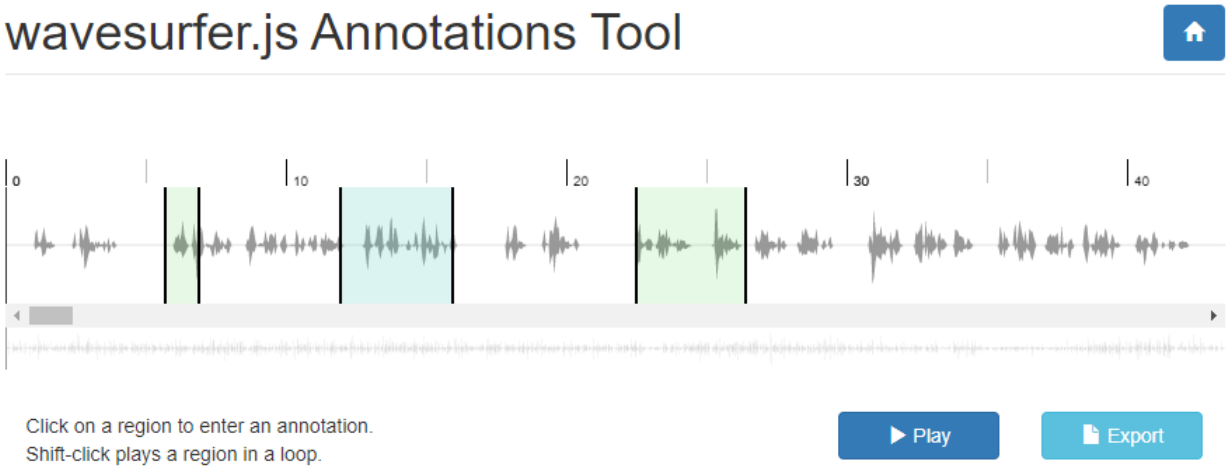
[3] Saiz, Matuszewski, and Goldszmidt, "Audio oriented UI components for the web platform", 2015

[4] Baker et al., "BioAcoustica: a free and open repository and analysis platform for bioacoustics", 2015

[5] Matuszewski, Schnell, and Goldszmidt, "Interactive Audiovisual Rendering of Recorded Audio and Related Data with the WavesJS Building Blocks", 2016

[6] Katspaugh, "wavesurfer.js", 2012

Existing online annotation tool[6]



- WaveSurfer is a general purpose JavaScript.
- It has been used via crowdsourcing in many applications include – 1) combining audios[3], 2) voice activity detection[4], and 3) audio rendering[5].
- However, it is not correction friendly.
  - Can be used for new annotation but may not be for correction.
  - Continuous zoom control.

[3] Saiz, Matuszewski, and Goldszmidt, "Audio oriented UI components for the web platform", 2015

[4] Baker et al., "BioAcoustica: a free and open repository and analysis platform for bioacoustics", 2015

[5] Matuszewski, Schnell, and Goldszmidt, "Interactive Audiovisual Rendering of Recorded Audio and Related Data with the WavesJS Building Blocks", 2016

[6] Katspaugh, "wavesurfer.js", 2012

Functionality of SPIRE-ABC

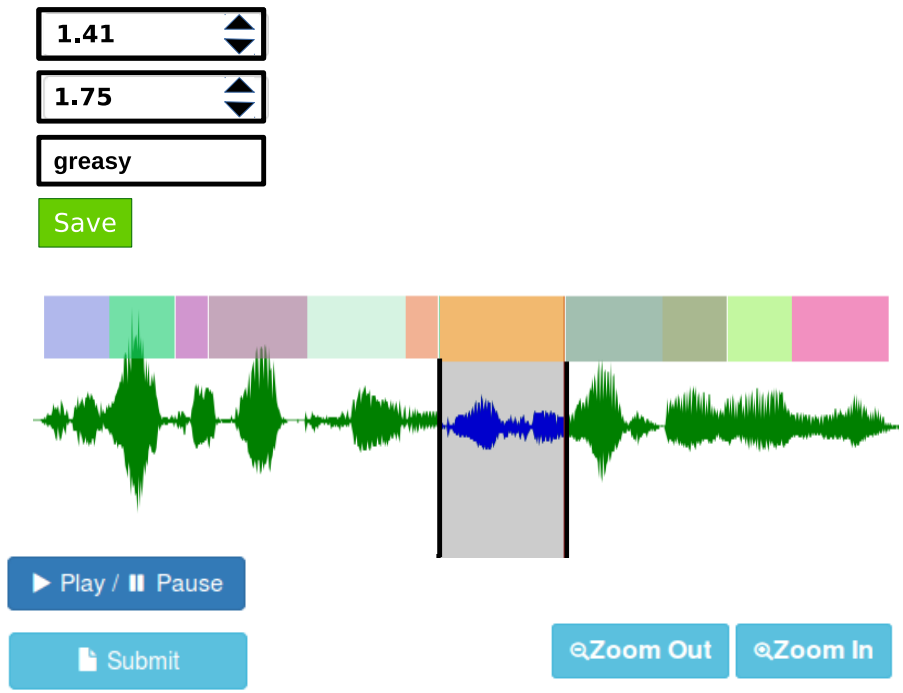


Figure: Annotation Interface of the SPIRE-ABC with an exemplary speech segment of *"she had your dark suit in greasy wash water all the year"*.

Functionality of SPIRE-ABC

- Two types of regions markings
  - Reference AU regions
  - Highlighted region for ABC (HR-ABC)

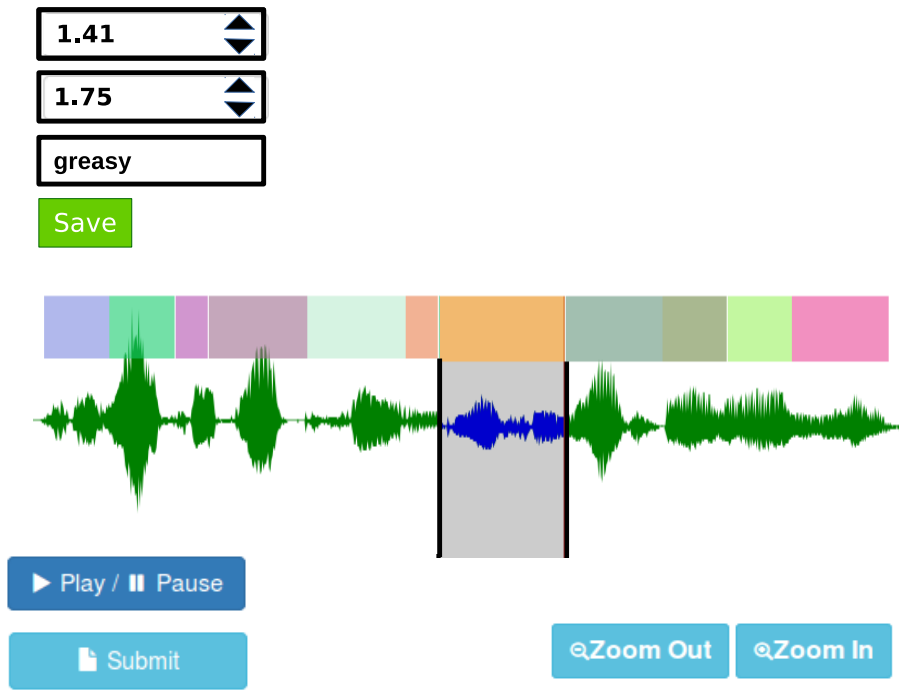


Figure: Annotation Interface of the SPIRE-ABC with an exemplary speech segment of *"she had your dark suit in greasy wash water all the year"*.

Functionality of SPIRE-ABC

- Two types of regions markings
  - Reference AU regions
  - Highlighted region for ABC (HR-ABC)
- Controls only specific to HR-ABC
  - Play the audio segment in HR-ABC
  - Zoom
  - Resizing by dragging the boundaries

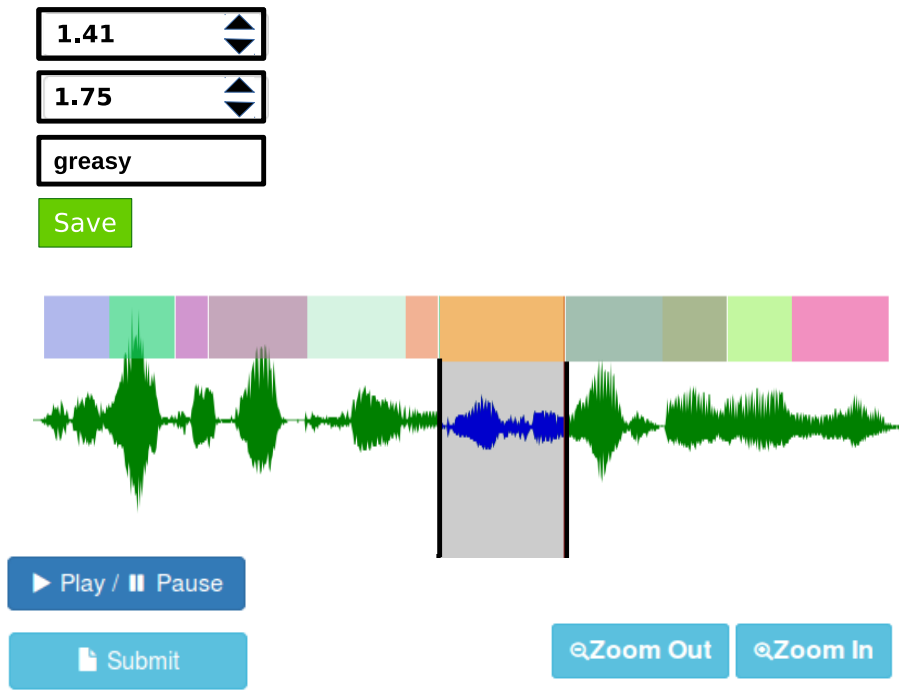


Figure: Annotation Interface of the SPIRE-ABC with an exemplary speech segment of *"she had your dark suit in greasy wash water all the year"*.

Functionality of SPIRE-ABC

- Two types of regions markings
  - Reference AU regions
  - Highlighted region for ABC (HR-ABC)
- Controls only specific to HR-ABC
  - Play the audio segment in HR-ABC
  - Zoom
  - Resizing by dragging the boundaries
- With save, selected reference regions (SRR-ABC) are updated based on HR-ABC

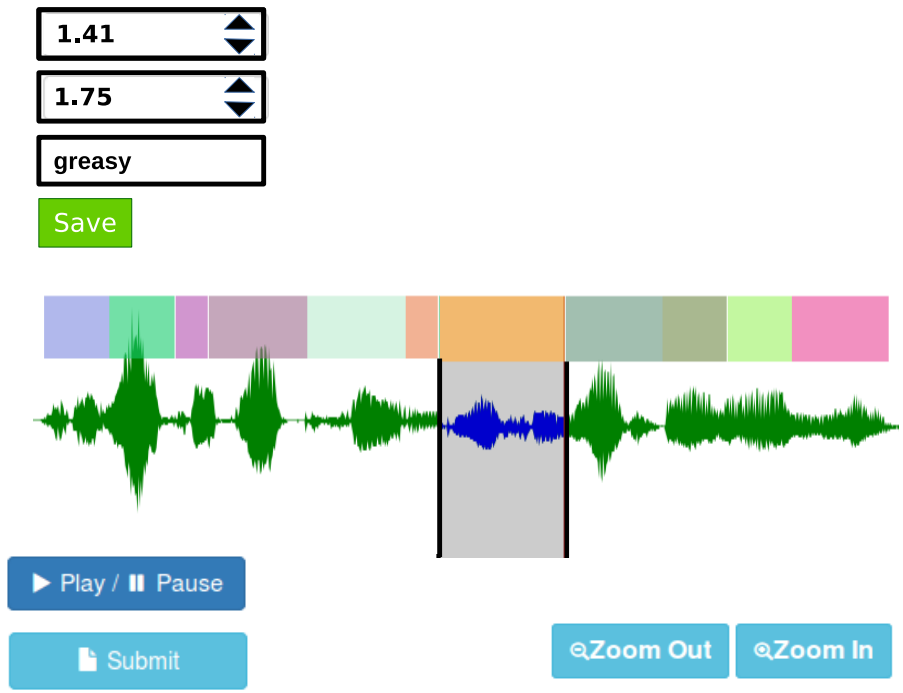
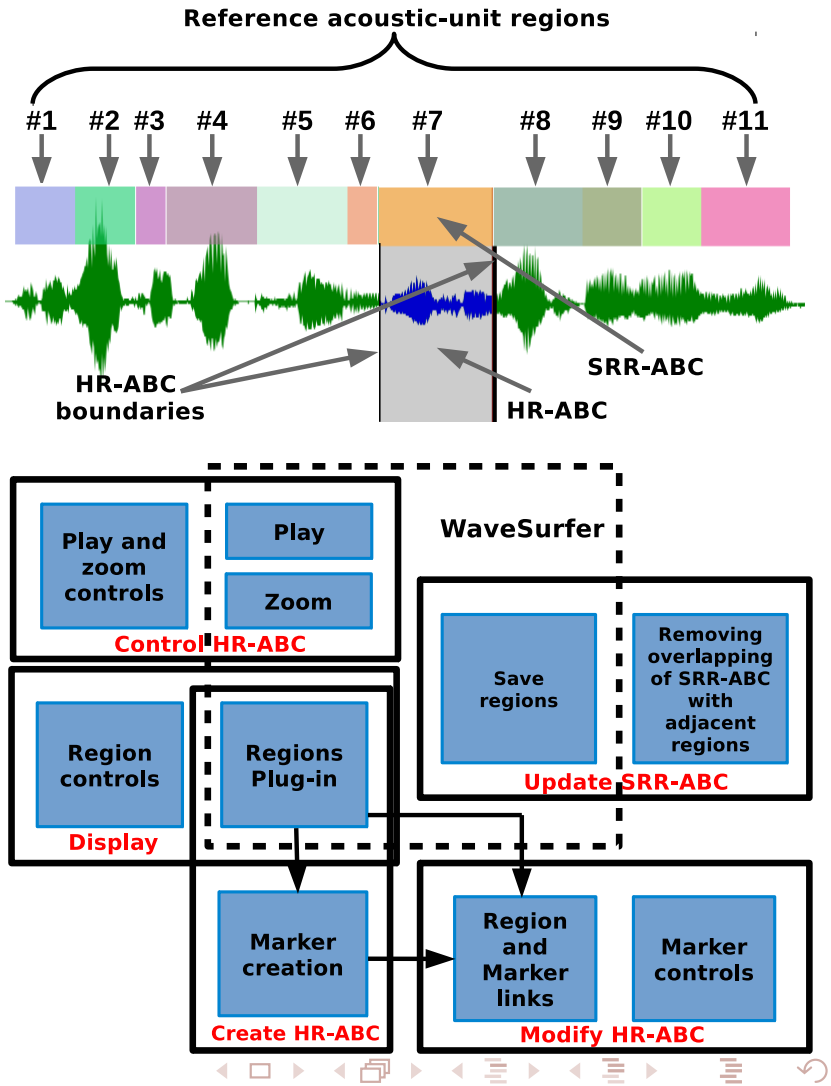


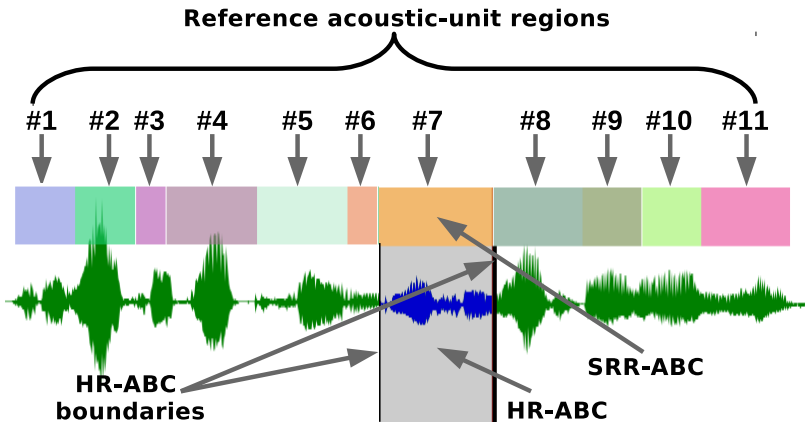
Figure: Annotation Interface of the SPIRE-ABC with an exemplary speech segment of *"she had your dark suit in greasy wash water all the year"*.

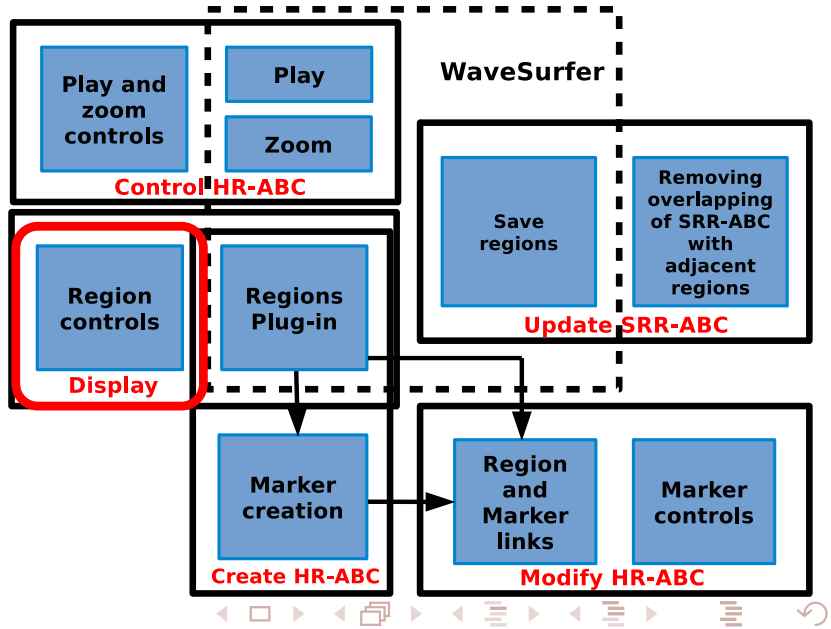
Proposed additional functionalities



Proposed additional functionalities

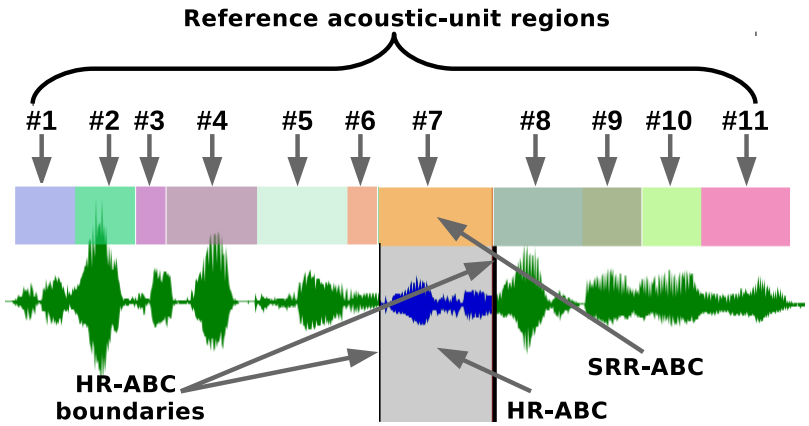
- Display SRR-ABC with only play option

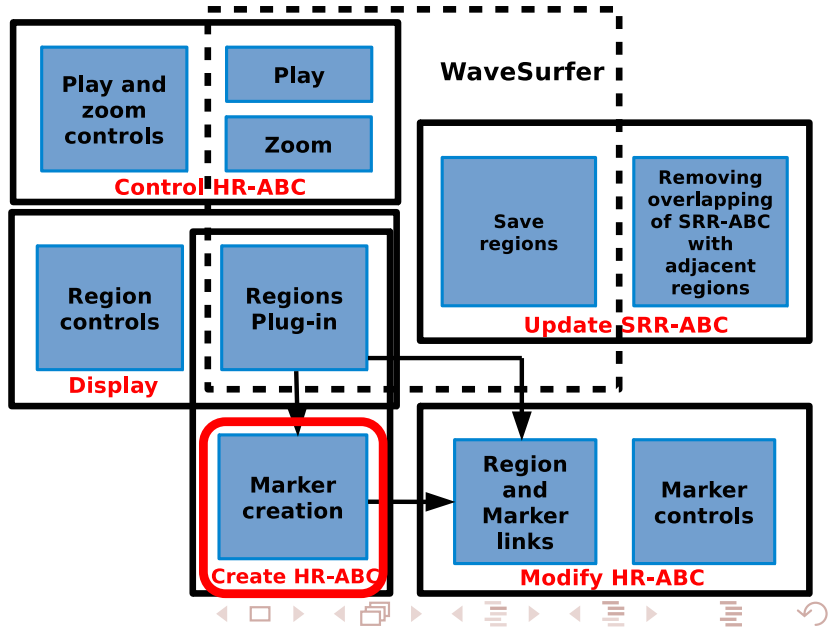




Proposed additional functionalities

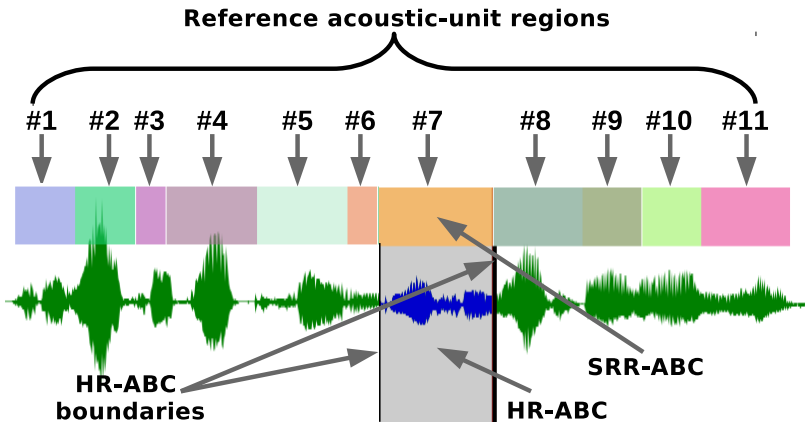
- Display SRR-ABC with only play option
- Create HR-ABC on mouse click on SRR-ABC with play, resize and move controls.

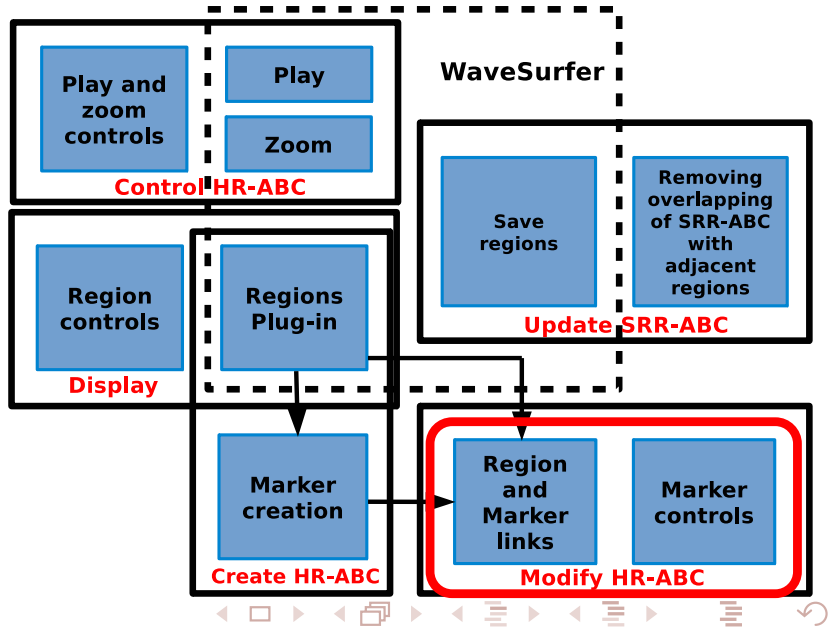




Proposed additional functionalities

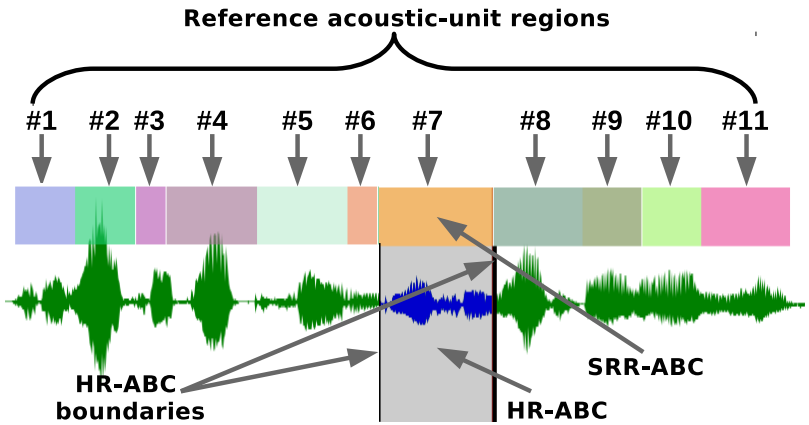
- Display SRR-ABC with only play option
- Create HR-ABC on mouse click on SRR-ABC with play, resize and move controls.
- Modify HR-ABC
  - Link the SRR-ABC and HR-ABC

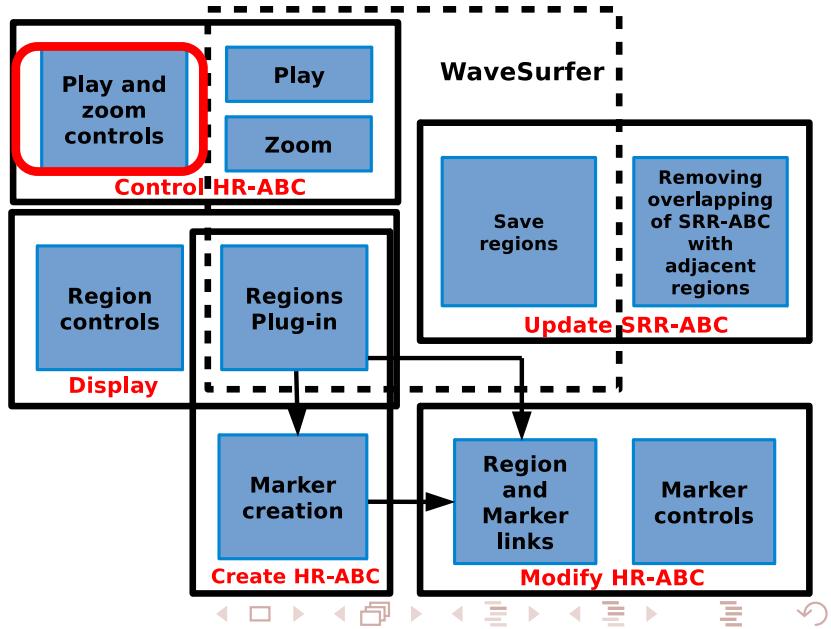




Proposed additional functionalities

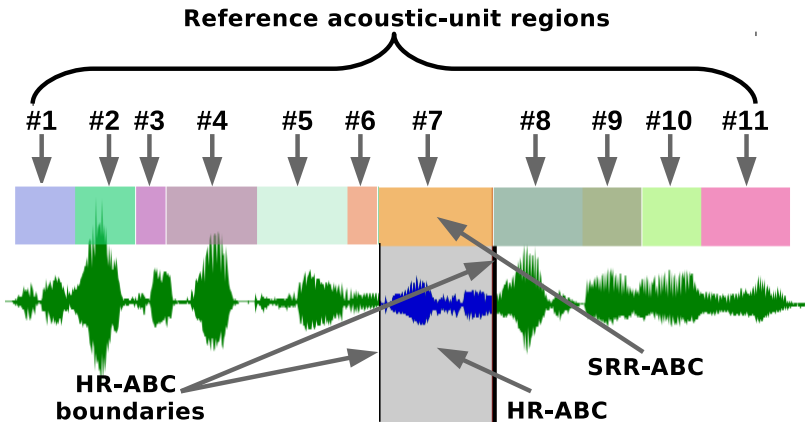
- Display SRR-ABC with only play option
- Create HR-ABC on mouse click on SRR-ABC with play, resize and move controls.
- Modify HR-ABC
  - Link the SRR-ABC and HR-ABC
- Control HR-ABC
  - Discrete zoom levels – $\frac{1}{2}$x, $\frac{1}{4}$x, $\frac{1}{8}$x, and $\frac{1}{16}$x,

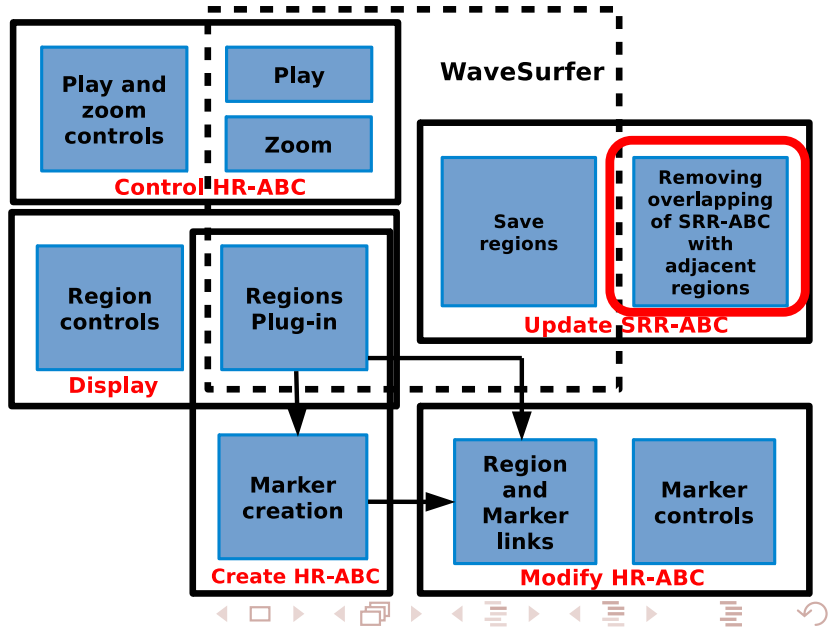




Proposed additional functionalities

- Display SRR-ABC with only play option
- Create HR-ABC on mouse click on SRR-ABC with play, resize and move controls.
- Modify HR-ABC
  - Link the SRR-ABC and HR-ABC
- Control HR-ABC
  - Discrete zoom levels – $\frac{1}{2}$x, $\frac{1}{4}$x, $\frac{1}{8}$x, and $\frac{1}{16}$x,
- Update SRR-ABC with save option





Experimental setup

Objective measures

- Mean absolute difference (MAD) between the ground truth and the corrected AU boundaries.
- Correct alignment rate (CAR): The percentage of AU boundaries that fall within a tolerance of 40ms from the ground truth AU boundaries.
- Overlap rate (OVR): The amount of overlap between the corrected and ground truth segments for all AUs.

Experimental setup

Objective measures

- Mean absolute difference (MAD) between the ground truth and the corrected AU boundaries.
- Correct alignment rate (CAR): The percentage of AU boundaries that fall within a tolerance of 40ms from the ground truth AU boundaries.
- Overlap rate (OVR): The amount of overlap between the corrected and ground truth segments for all AUs.

- 30 utterances from TIMIT data are considered, for which, ground-truth AU boundaries are available.

Experimental setup

Objective measures

- Mean absolute difference (MAD) between the ground truth and the corrected AU boundaries.
- Correct alignment rate (CAR): The percentage of AU boundaries that fall within a tolerance of 40ms from the ground truth AU boundaries.
- Overlap rate (OVR): The amount of overlap between the corrected and ground truth segments for all AUs.

- 30 utterances from TIMIT data are considered, for which, ground-truth AU boundaries are available.
- AU segments – syllable and words, obtained with fisher English and TIMIT data. Total: FE_S; TIMIT_S; FE_W; TIMIT_W.

Experimental setup

Objective measures

- Mean absolute difference (MAD) between the ground truth and the corrected AU boundaries.
- Correct alignment rate (CAR): The percentage of AU boundaries that fall within a tolerance of 40ms from the ground truth AU boundaries.
- Overlap rate (OVR): The amount of overlap between the corrected and ground truth segments for all AUs.

- 30 utterances from TIMIT data are considered, for which, ground-truth AU boundaries are available.
- AU segments – syllable and words, obtained with fisher English and TIMIT data. Total: FE_S; TIMIT_S; FE_W; TIMIT_W.
- Annotators – Experienced (EA), Inexperienced (IEA), naive (NA).

Experimental setup

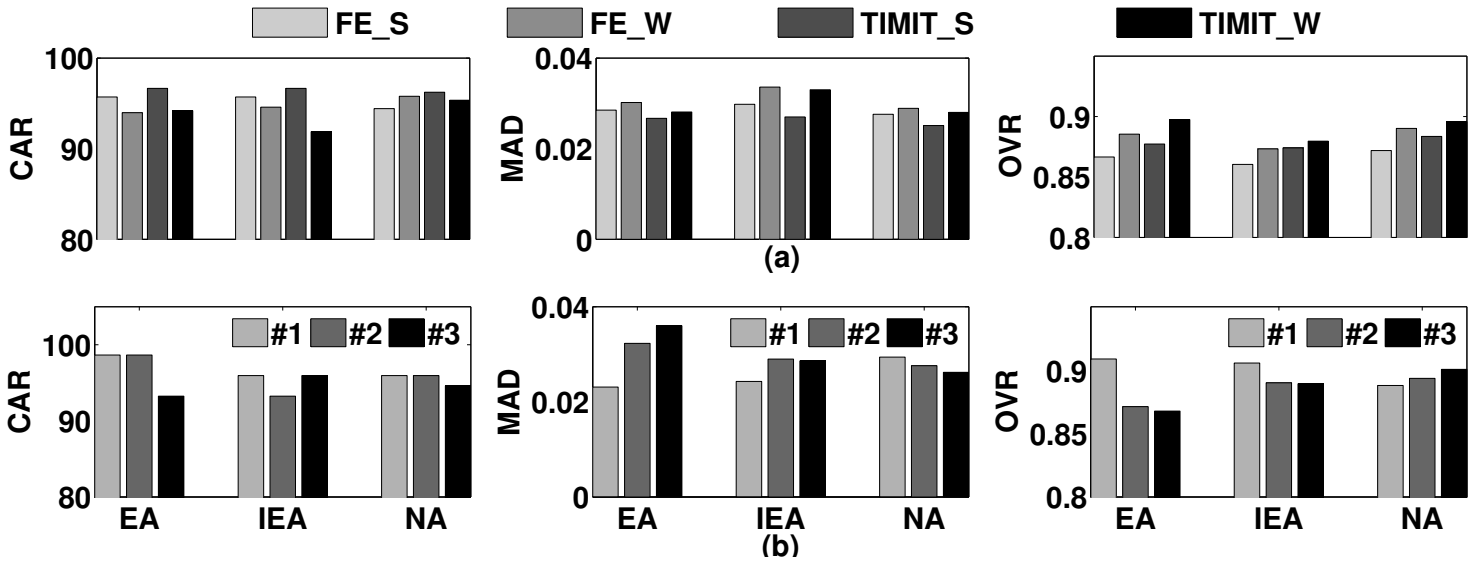
Objective measures

- Mean absolute difference (MAD) between the ground truth and the corrected AU boundaries.
- Correct alignment rate (CAR): The percentage of AU boundaries that fall within a tolerance of 40ms from the ground truth AU boundaries.
- Overlap rate (OVR): The amount of overlap between the corrected and ground truth segments for all AUs.

- 30 utterances from TIMIT data are considered, for which, ground-truth AU boundaries are available.
- AU segments – syllable and words, obtained with fisher English and TIMIT data. Total: FE_S; TIMIT_S; FE_W; TIMIT_W.
- Annotators – Experienced (EA), Inexperienced (IEA), naive (NA).

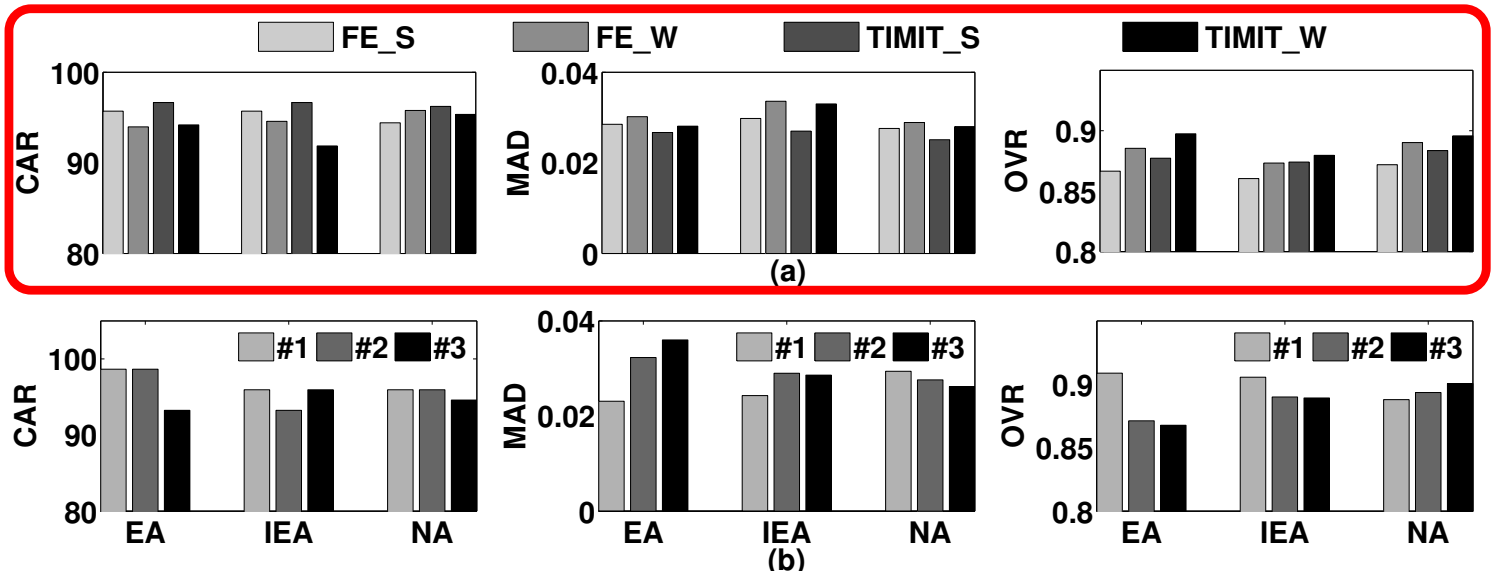
	FE_S	FE_W	TIMIT_S	TIMIT_W	Common	Total
EA#1, IEA#1, NA#1	10	10	10	10	12	52
EA#2, IEA#2, NA#2	10	10	10	10	12	52
EA#3, IEA#3, NA#3	10	10	10	10	12	52

Results



	FE_S	FE_W	TIMIT_S	TIMIT_W
CAR	83.02	78.92	86.72	82.35
MAD	0.0465	0.0518	0.0352	0.0394
OVR	0.7927	0.8120	0.8257	0.8398

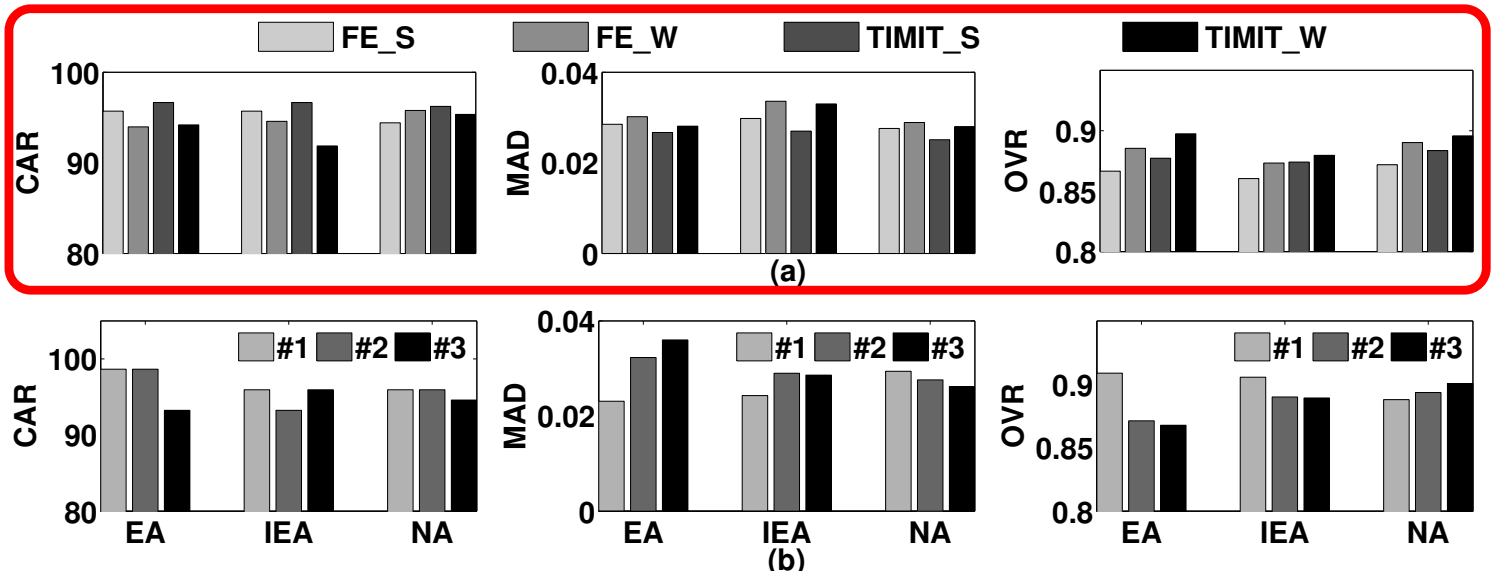
Results



	FE_S	FE_W	TIMIT_S	TIMIT_W
CAR	83.02	78.92	86.72	82.35
MAD	0.0465	0.0518	0.0352	0.0394
OVR	0.7927	0.8120	0.8257	0.8398

- After manual correction, all the three type annotators has shown improved performance.

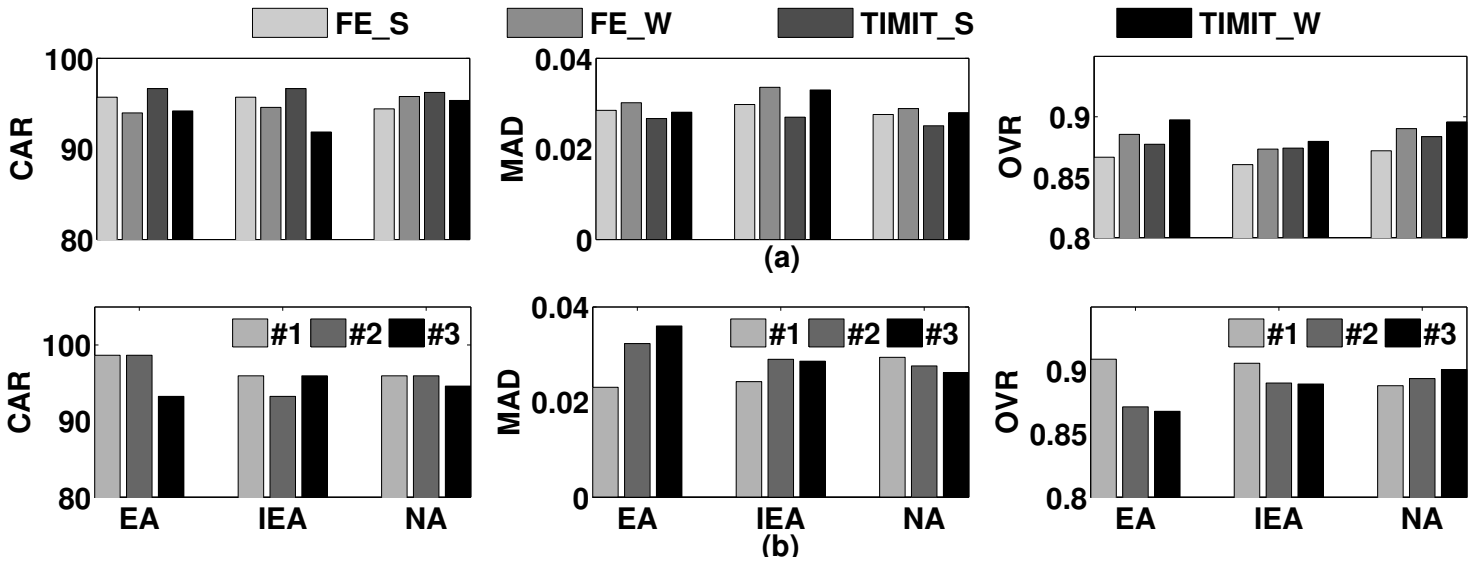
Results



	FE_S	FE_W	TIMIT_S	TIMIT_W
CAR	83.02	78.92	86.72	82.35
MAD	0.0465	0.0518	0.0352	0.0394
OVR	0.7927	0.8120	0.8257	0.8398

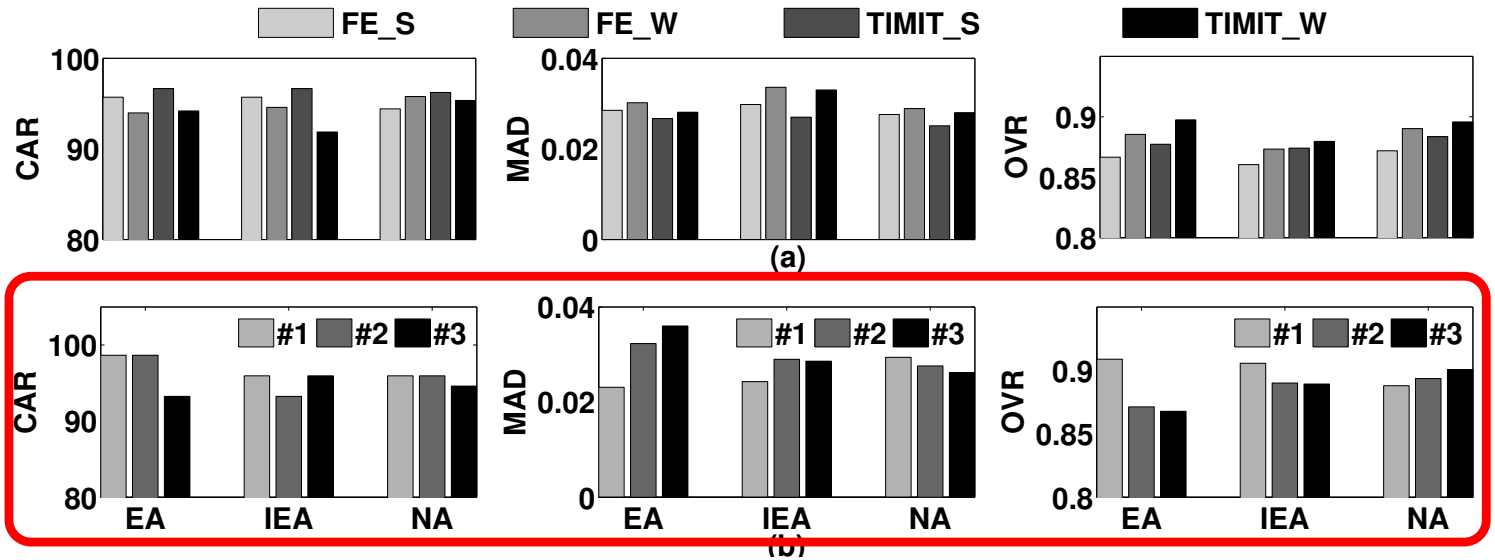
- After manual correction, all the three type annotators has shown improved performance.
- The performance measures obtained by NAs are not significantly different from those by EAs and IEAs.

Results



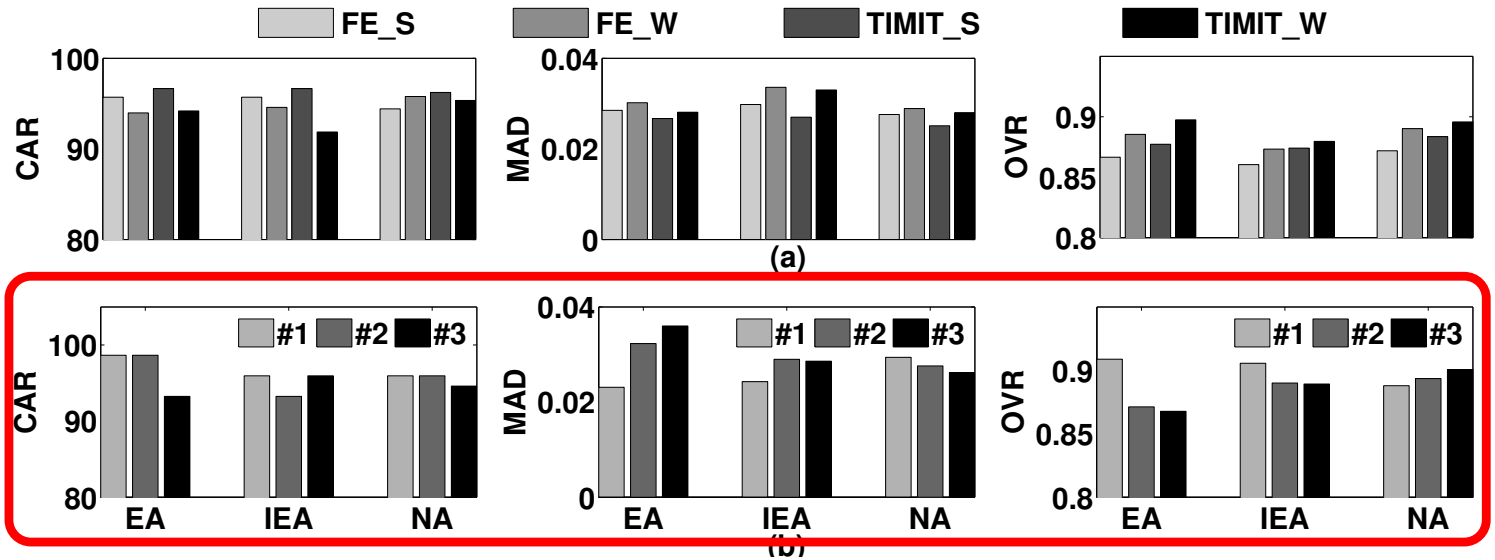
- NAs have higher CAR than both EAs and IEAs for TIMIT_W setup.

Results



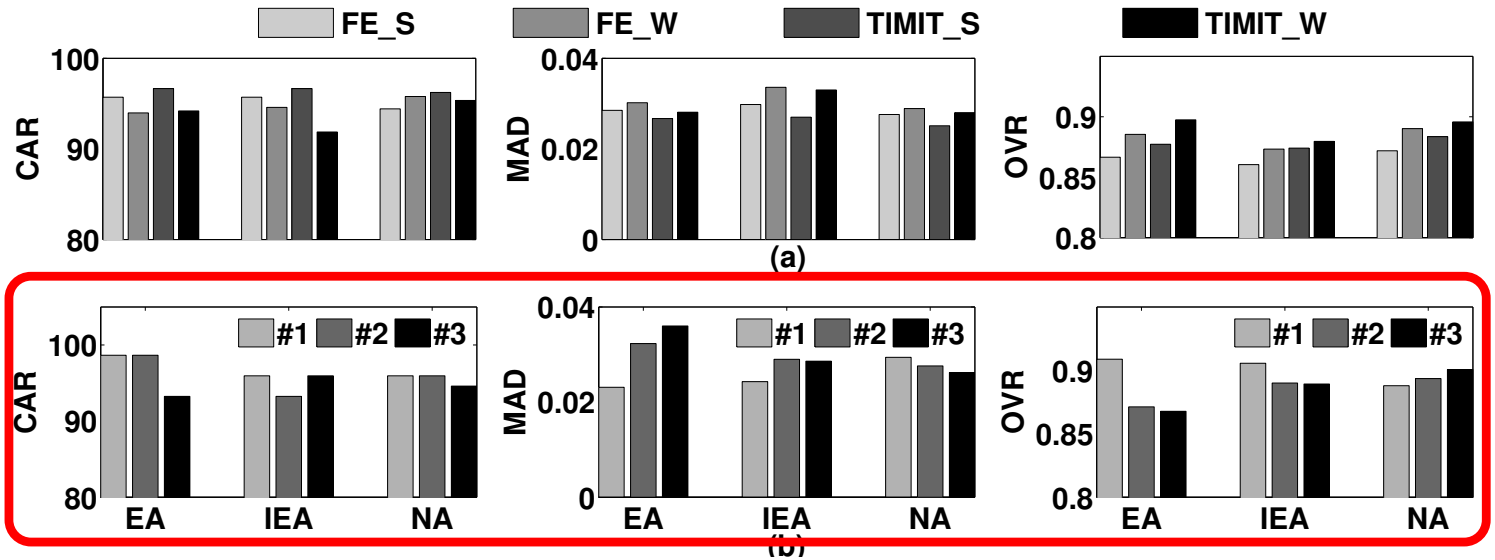
- NAs have higher CAR than both EAs and IEAs for TIMIT_W setup.

Results



- NAs have higher CAR than both EAs and IEAs for TIMIT_W setup.
- On the common set, EA#1 shows better performance across all performance measures over IEAs and NAs.

Results



- NAs have higher CAR than both EAs and IEAs for TIMIT_W setup.
- On the common set, EA#1 shows better performance across all performance measures over IEAs and NAs.
- However, interestingly, the EA#3 has lower performance among all EAs and across both the IEAs and NAs.

Conclusion

- This work presents SPIRE-ABC that helps in correcting errors in noisy acoustic-unit boundaries using web interface via crowdsourcing.

Conclusion

- This work presents SPIRE-ABC that helps in correcting errors in noisy acoustic-unit boundaries using web interface via crowdsourcing.
- This is developed by creating additional functional modules as well as modifying the existing functional modules in the WaveSurfer.

Conclusion

- This work presents SPIRE-ABC that helps in correcting errors in noisy acoustic-unit boundaries using web interface via crowdsourcing.
- This is developed by creating additional functional modules as well as modifying the existing functional modules in the WaveSurfer.
- Experiments on TIMIT corpus have shown improvements in the AU boundaries after manual correction irrespective of annotators type.

Conclusion

- This work presents SPIRE-ABC that helps in correcting errors in noisy acoustic-unit boundaries using web interface via crowdsourcing.
- This is developed by creating additional functional modules as well as modifying the existing functional modules in the WaveSurfer.
- Experiments on TIMIT corpus have shown improvements in the AU boundaries after manual correction irrespective of annotators type.
- Further works are required for adding all reference acoustic-unit transcriptions.

THANK YOU

For more info:

http://spire.ee.iisc.ernet.in/spire-abc/