PESInet: Automatic Recognition of Italian Statements, Questions, and Exclamations With Neural Networks

Sonia Cenceschi
SUPSI University
Via Pobiette 11, Manno, Switzerland
sonia.cenceschi@supsi.ch

Licia Sbattella
Politecnico di Milano
P.za L. da Vinci 32, Milano, Italy
licia.sbattella@polimi.it

Roberto Tedesco
Politecnico di Milano
P.za L. da Vinci 32, Milano, Italy
roberto.tedesco@polimi.it

Davide Losio
Politecnico di Milano
P.za L. da Vinci 32, Milano
davide.losio@mail.polimi.it

Mauro Luchetti
Politecnico di Milano
P.za L. da Vinci 32, Milano, Italy
mauro.luchetti@mail.polimi.it

Abstract

PESInet is an Automatic Prosody Recognition system aiming at classifying Information Units as *Statement*, *Question* or *Exclamation*. PESInet adopts a modular architecture, with a master NN evaluating the results of two independent BLSTM NNs that work on audio and its transcription. PESInet has been trained with our own three-class, balanced corpus composed of about 1.5 million text phrases and 60 000 utterances of recited and spontaneous speech. PESInet reached an accuracy of 80% on three classes, and 91% on two classes (*Question* vs *Non-question*). Finally PESInet, compared against human listeners on a two-class test based on a different corpus, reached a better Accuracy (89% for PESInet, against 80% for human listeners).

1 Credits

The Prosody Extraction by Sound Interpreting network (PESInet) is part of the Lend Your Voice (LYV) project, which has been funded by the Polisocial Award[1] 2016, in collaboration with Fondazione Sequeri Esagramma[2].

2 Introduction

The goal of PESInet was to investigate whether clues derived from text could improve the recognition of simple prosodic forms in Information Units (IUs). In particular, we focused on *Statement*, *Question*, and *Exclamation* which are proposition's structures and are independent of the pragmatic function of the corresponding IU: each one can assume a large set of illocutionary acts, as explained into the Language into Act theory (L-AcT) described in Cresti (2014). An IU is composed of a textual realisation (i.e., a written phrase) and an acoustic realisation (an audio recording of a speaker uttering such a phrase), and conveys a specific informative intention (Austin, 1975; Cresti, 2000). We designed a modular model based on Neural Networks (NNs), able to highlight how much audio and text affected recognition accuracy. Moreover, to validate our results, we compared our NN model against human listeners, on a set of IUs that did not overlap with the corpus we used to train the model.

3 Background

The majority of studies on prosody regards the automatic recognition or detection of single prosodic clues (Ren et al., 2004; Jeon and Liu, 2009; Tamburini and Wagner, 2007; Taylor, 1993). Others, deal with the detection of phrase boundaries or prosodic phrases (Liu et al., 2006; Wightman and Ostendorf, 1991; Rosenberg, 2009). Just a few works, however, focus on *modality* detection. In the following we briefly introduce some of them. *Question* detection is investigated in Tang et al. (2016) using Recurrent Neural Networks (RNN), in the Mandarin language. Authors propose sev-

[1] http://www.polisocial.polimi.it

[2] https://www.esagramma.net

eral RNN and Bidirectional RNN (BRNN) models, trained on a simulated call-centre recordings consisting of just 2850 Question and 3142 Non-question IUs. The best result is an F_1 score of 85.5%.

The work described in Yuan and Jurafsky (2005) focuses on *Question* and *Statement* detection, from text and audio, for Chinese; authors investigate the influence of text in prosody comprehension, on a telephone corpus (with transcriptions). Their classifier achieves an error rate of 14.9% with respect to a 50% chance-level rate. Quang et al. (2007) use decision trees to automatically detect *Questions* in a small elicited French and Vietnamese corpora, leveraging both acoustic and lexical features (unigrams, bigrams, and presence of so-called "interrogative terms"). The best result is an F_1 of 80% for the Vietnamese language.

Finally, the work described in Li et al. (2016) combines Convolutional NNs (CNNs) and Bidirectional Long Short-Term Memory NNs (BLSTM) to extract textual and acoustic features for recognising stances (Affirmative, Neutral, Negative opinions) in the Mandarin language. It exploits a small, manually-tagged corpus of four debate videos (1254 IUs). Combining both audio and text this system reaches an Accuracy of 90.3%.

None of the works mentioned above is perfectly comparable with ours and, on the other hand, all of them are based on ah-hoc corpora (as we did). This makes impossible to compare the results we obtained against other approaches. We, however, validated our results comparing our model against human listeners.

4 The corpus

Our own corpus is composed of eBooks, EPUB3 audio-books (an EPUB3 audio-book contains both text and audio recording, time-aligned at the level of sentence), and the LIT/DIA-LIT corpus (Biffi, 1976; Buroni, 2009), which contains audio recordings of Italian TV shows, with transcriptions.

From eBooks, the textual part of EPUB3 audio-books, and transcriptions of LIT/DIA-LIT we extracted about 1.5 million sentences, balanced on the three target classes: *Statement*, *Question*, and *Exclamation*.

From LIT/DIA-LIT audio recordings and the audio part of EPUB3 audio-books, we collected about 60 000 utterances (again, balanced on the three target classes). Both sentences and utterances were tagged with the correct class, leveraging the punctuation marks we found in text/transcriptions. Of course, we removed such punctuation marks from the textual part of the corpus. Moreover, we discarded all the sentences containing a sub-phrase or other complex syntactic structures. In doing so we aimed at retaining plain simple examples of statements, questions, and exclamations.

We are aware that leveraging punctuation marks for tagging sentences can lead to confounds, as exclamation marks is also used for Vocatives and Orders, while the full stop is also used for Orders. Anyway, it was simply not possible to manually review the text collection and manually solve the problem. Thus, we assume our corpus is affected by a small amount of noise (in other words, we assume Exclamations and Statements are way more frequent than Vocatives and Orders).

Notice that the question marks might be used for different question typologies (rhetorical, information-seeking, confirmation-seeking, biassed), and that question could be further partitioned into open questions, polar questions, etc. Thus, the question mark is used to tag sentences with wildly divergent phonetic forms. This is not, however, a blocking issue: it only makes harder for the classifier to learn the input/output correlation. In particular, this is one of the reasons that lead us to the idea of leveraging text to improve the classification of IUs.

Summing up, we built three corpora:

- **ACorpus**: audio corpus composed of about 60 000 .wav labelled samples.
- **TCorpus**: textual corpus composed of about 1.5 million .txt labelled samples.
- **MCorpus**: mixed corpus composed of all the ACorpus files, with their transcriptions (from the TCorpus); about 60 000 labelled samples.

5 Features extraction

From acoustic and textual samples we derived a set of features that our NNs leveraged for training and recognition.

5.1 Acoustic features

With a sample rate of 44.1 kHz, we adopted a window of 2048 samples with a hop-size of 1024 samples (i.e., every 23 ms a new vector of acoustic features is produced). Notice that our window is larger than the one usually adopted by ASRs; in fact, we are not interested in phone recognition and, on the other hand, prosody phenomena appear in larger temporal scale than the one involving individual phones.

We tried several window sizes, and several acoustic features; in particular we experimented with different combinations of Cepstrum coefficients. At the end, we come up with the following 129 acoustic features, normalised (to minimise dependency on speakers and recording settings) and calculated by means of Praat (Boersma and others, 2001), as they provided the best results:

- pitch value, with its delta and delta-delta
- energy, with its delta and delta-delta
- the first 40 Cepstrum coefficients, with their deltas and delta-deltas
- energy of such 40 Cepstrum coefficients (as MFCC defines), with its delta and delta-delta

Notice that we did not adopt a true "deep" architecture, as features were not "discovered" by the network. The field of audio analysis already provides a huge set of well-known, informative features; thus, in our opinion, there is no point in let the network approximating them. Moreover, precalculated features permit to simplify the network. Summing up, each utterance was transformed into an array that contains a column of 129 real numbers every 23ms.

5.2 Textual features

To feed the model with textual samples we used the usual word embedding technique, which represents the vocabulary in a continuous vector space of 300 dimensions (Sahlgren, 2008). In particular we adopted Italian Word Embeddings, a pretrained model of 700 000 words based on GloVe (Pennington et al., 2014).

Summing up, each sentence was transformed into an array that contains a column of 300 real numbers for each token. Notice that punctuation marks were discarded and no lemmalisation was applied.

6 Architecture

PESInet is composed of three different NNs:

1. Audio-based NN
2. Text-based NN
3. Master NN combining the prediction of the two preceding NNs

We developed two NN architectures: for Audio-based and Text-based NNs, and for Master NN.

6.1 The convolutional block

Acoustic and textual features defined in Section 5 generated low-level pieces of information, looking at very local phenomena. For considering higher-level phenomena, both the Audio-based and the Text-based NNs relied on the same architecture, leveraging an initial multi-layer convolutional block.

A convolutional layer is composed of several kernels with a predefined width, which "scan" the input array. Each kernel, after the training phase, specialises in finding certain *patterns* in the input sequence. The network learns "high level" features (i.e., common prosody contours, for the Audio-based NN, or particular word sequences for the Text-based NN) from our low-level feature set.

Features related to prosody unfold along different time extents (Cutugno et al., 2005): we found dependencies both in short and long time periods. So the idea was to use different kernel widths, in order to allow the network to consider different pattern lengths. The hint to adopt this technique come from various papers (Sbattella et al., 2014; Gussenhoven, 2008; Büring and others, 2009), which thoroughly analysed the idea of simultaneously analysing the input at different temporal granularities with the use of differently-sized kernels.

In particular, our convolutional block is composed of three layers, which "scan" at three different temporal granularity levels. In general, if s is the stride adopted for kernels at any temporal granularity level and d_i is the kernel height at the i-th temporal granularity level, the kernel height at the $(i + 1)$-th temporal granularity level is $d_{i+1} = d_i + s$; see Figure 1, for a simplified example with two levels. Stride is chosen so that, after each shift of the filter, the kernel will include a small subset of the previously analysed input.

Available at: http://hlt.isti.cnr.it/wordembeddings/

Finally, padding is applied to the input sequence, so that the shorter kernels (and, by construction, all the other, longer kernels) fit the sequence length.

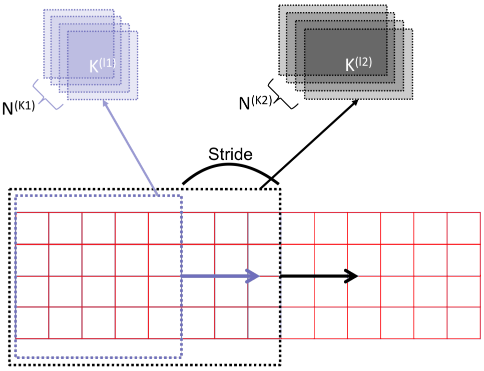

Figure 1: Kernels $K^{(l1)}$ and $K^{(l2)}$ at two different temporal granularity levels.

Being the kernels of different heights, they will cause the outputs to have different dimension as well, relatively to the layers they come from. These dimensions are adjusted in the following layer of the network. Figure 2 shows a simplified schema with two differently-sized kernel groups.

6.2 Audio-based and Text-based NNs

Both the Audio-based and the Text-based NNs relied on a multi-layer network. The general architecture is composed of three BLSTM layers on top of the convolutional block. We connected the first convolutional layer to the first BLSTM layer; then, the second convolutional layer is connected, together with the output from the first BLSTM, to the second BLSTM layer; finally, the third convolutional layer is connected, together with the output of the second BLSTM layer, to the third BLSTM layer. Figure 3 shows the way in which the convolutional block is used.

The Softmax layer shown in the Figure 3 is used during the training phase and then removed, as the Text-based and Audio-based NNs are combined together with the Master NN.

6.3 Master NN and PESInet

The Master NN is composed of a fully-connected layer, and a Softmax layer. PESInet, the resulting network, is shown in Figure 4. Notice that PESInet is supposed to works on utterances, while the text is generated by means of an ASR; in fact, this is the setting we expect to be adopted during actual usage of PESInet. Our corpus, conversely, was based on human-generated text; we are aware that in doing so we did not consider the errors due to the ASR and, as a consequence, overestimated the figures obtained during the training/validation procedure. The rationale was highlighting the contribution of text-related features to the recognition of prosodic forms, and thus we decided to avoid the “noise” introduced by ASR-related errors.

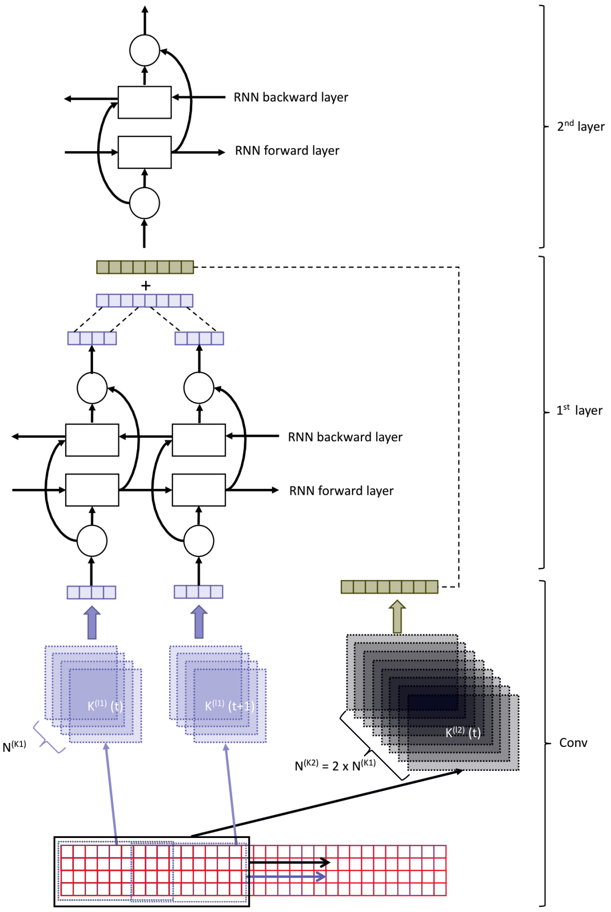

Figure 2: Convolution with two kernel sizes (i.e., two temporal granularity levels).

As a final remark on the ASR, notice that it is supposed to not add any punctuation mark to the transcription it generates.

7 Training

The architecture was implemented, trained, and tested using the TensorFlow library along with Python 3.6. The code itself was run on a machine equipped with 32GB of RAM, a Xeon Intel processor and a Nvidia Titan X (Pascal) GPU. During training, we adopted the early stopping (using Accuracy as reference index); moreover, to improve

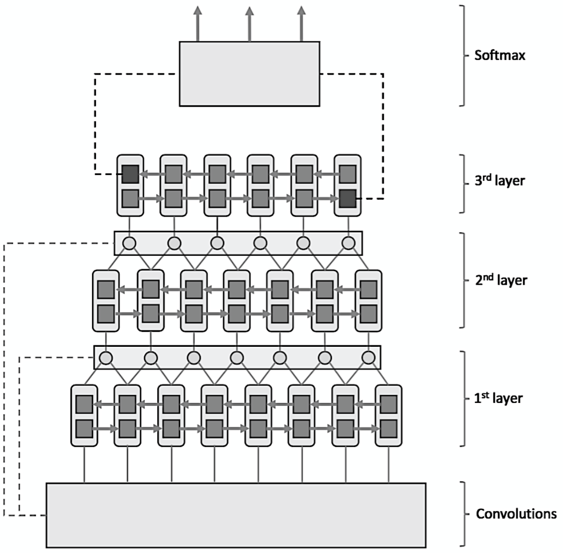

Figure 3: Structure of the Text-based and Audio-based NNs.

the learning effectiveness, we used the variational drop-out on recurrent layers. We started training, independently, the Audio-based and the Text-based NNs, on 80% of their respective corpora: ACorpus and TCorpus. Then, once removed the final Softmax layer from them, these NNs where attached to the Master NN, and a further training –involving 80% of the MCorpus– was performed on PESInet. In particular, we investigated three approaches:

1. Allowing PESInet to train only the Master NN weights (all the others remain fixed).
2. Allowing PESInet to change all its internal weights (also those already trained).
3. Training PESInet from scratch, skipping training of Audio-based and Text-based NNs.

8 Evaluation

Validation was performed using 20% of the corpus. We experimented with several feature combinations, hyperparameter values, and network structures, before reaching the final models.

The Audio-based and Text-based NNs gave the following Accuracies: 0.68 and 0.79. It's interesting that the Text-based NN gave a better Accuracy than the Audio-based NN. This was surprising, as, after all, prosody is an acoustic phenomenon. Nevertheless, data seem to show that the words composing the utterance are indeed a good predictor of prosody. Moreover, considering that ACorpus was much smaller than TCorpus, the surprisingly low results of Audio-based NN can be explained.

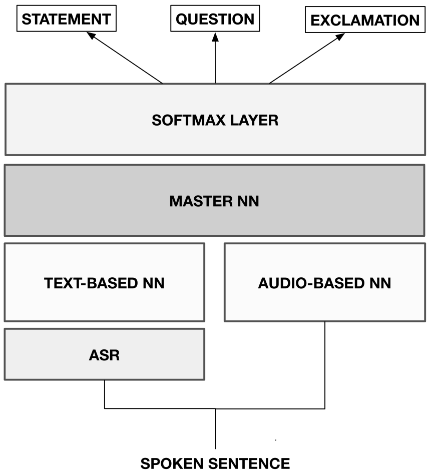

Figure 4: PESInet structure.

		Predicted		
		Stat.	Excl.	Quest.
True	Stat.	1366	234	155
	Excl.	285	1068	316
	Quest.	216	484	1130

Table 1: Confusion matrix for Audio-based NN.

Table 1 and Table 2 show the confusion matrices for the two NNs. It's interesting to notice that Audio-based NN predicted Statements much better than the other two classes, while Text-based NN was also very good in recognising Questions.

About PESInet, Table 3 shows that the approach 2 obtained, as expected, the best results. As the confusion matrix of Table 4 shows, audio and text cooperated to improve recognition of all the three classes.

As a further experiment, we trained and tested PESInet on two classes: *Question* vs *Non-question*, adapting the same PESInet architecture to handle 2 classes. The corpus tags

		Predicted		
		Stat.	Excl.	Quest.
True	Stat.	48 478	7233	3358
	Excl.	8786	43 887	6064
	Quest.	4494	5905	48 495

Table 2: Confusion matrix for Text-based NN.

Trained NN	**PT**	**F_1**	**Loss**	**Acc.**
1. Master NN	yes	0.79	0.55	0.77
2. PESInet	yes	**0.80**	**0.49**	**0.80**
3. PESInet	no	0.80	0.55	0.78

Table 3: Results for PESInet. PT: Pre-training Text-based and Audio-based NNs.

		Predicted		
		Stat.	Excl.	Quest.
True	Stat.	1444	205	106
	Excl.	222	1242	205
	Quest.	92	215	1523

Table 4: Confusion matrix for PESInet.

Trained NN	**PT**	**F_1**	**Loss**	**Acc.**
2. PESInet	yes	0.91	0.39	0.91

Table 5: Results for PESInet, two classes.

{*Exclamation*, *Statement*} were rewritten as *Non-question*, and we randomly extracted a number of *Non-Question* samples equals to the *Question* samples. Then, we used 90% of such dataset for training and 10% for testing. Accuracy reached 91% (Table 5).

8.1 PESInet against human listeners

Finally, to validate the results we obtained, we conducted a perceptive experiments with 302 Italian speakers (Cenceschi et al., 2018b; Cenceschi et al., 2018a). The aim of the experiment was to understand the role of acoustic clues and textual clues in the perception of various prosodic forms.

The experiment was divided into several tests; each test was about a specific prosodic form: users were asked to listen a set of IUs and select which of them carried the expected prosodic form. In that experiment we used an ad-hoc audio/textual corpus called SI-CALLIOPE, where 14 professional actors spoke a set of 139 sentences, for a total of 1946 IUs. Notice that SI-CALLIOPE did not share anything, in terms of sentences and speakers, with corpora we used to train PESInet.

In particular, for the *Question/Non-question* test, each user listened to a set of audios randomly extracted from 714 question IUs and 1232 non-question IUs. The average accuracy was 80% (std. dev.: 7.24%).

Running the two-class version of PESInet on the same test, we got an Accuracy of 89%.

We argue that this surprisingly good Accuracy for our NN (or surprisingly bad Accuracy for human listeners) could be caused by *de-contextualisation*: in the experiment each IU was given in isolation, without any dialogue context; probably, listeners were more affected by that lacking of context than our NN. Anyway, this is just a hypothesis that should be investigated and deepened with further experiments, as the comparison could be tainted by a large number of other confounds, such as the non ecological nature of the task and the stratification of the repertoire of Italian speakers.

9 Conclusions and discussion

PESInet got an Accuracy of 80% on three classes and and 91% on two classes. Moreover, PESInet reached very good results when compared to human listeners on a totally different corpus. Although this human/NN comparison should be taken with a grain of salt, we believe that it is a hint that the network works well and the results are truly promising. As a future work, more recordings should be added to ACorpus and MCorpus to improve the performance of the Audio-based NN and, as a consequence, of the whole PESInet.

Currently, we are working for cleaning the code and streamlining the training procedure, as we plan to release the code.

References

John Langshaw Austin. 1975. *How to do things with words*, volume 88. Oxford university press.

Marco Biffi. 1976. Il lit–lessico italiano televisivo: l'italiano televisivo in rete. *L'italiano televisivo: 1976-2006. Atti del convegno–Milano, 15-16 giugno 2009*, pages 35–69.

Paul Boersma et al. 2001. Praat, a system for doing phonetics by computer. *Glot international*, 5(9/10):341–345.

Daniel Büring et al. 2009. Towards a typology of focus realization. *Information structure*, pages 177–205.

Edoardo Buroni. 2009. La voce del telegiornale. aspetti prosodici del parlato telegiornalistico italiano in chiave diacronica. l'italiano televisivo 1976–2006. *Atti del Convegno, Milano*, pages 15–16.

Sonia Cenceschi, Licia Sbattella, and Roberto Tedesco. 2018a. Towards automatic recognition of prosody. In *Proc. 9th International Conference on Speech Prosody 2018*, pages 319–323.

Sonia Cenceschi, Licia Sbattella, and Roberto Tedesco. 2018b. Verso il riconoscimento automatico della prosodia. *STUDI AISV*, pages 433–440.

Emanuela Cresti. 2000. *Corpus di italiano parlato: Introduzione*, volume 1. Accademia della Crusca.

F. Cutugno, G. Coro, and M. Petrillo. 2005. Multigranular scale speech recognizers: Technological and cognitive view. In Springer, editor, *Congress of the Italian Association for Artificial Intelligence*, 227–330, Berlin.

Carlos Gussenhoven. 2008. Types of focus in english. In *Topic and focus*, pages 83–100. Springer.

Je Hun Jeon and Yang Liu. 2009. Automatic prosodic events detection using syllable-based acoustic and syntactic features. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4565–4568. IEEE.

Linchuan Li, Zhiyong Wu, Mingxing Xu, Helen M Meng, and Lianhong Cai. 2016. Combining cnn and blstm to extract textual and acoustic features for recognizing stances in mandarin ideological debate competition. In *INTERSPEECH*, pages 1392–1396.

Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. 2006. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *IEEE Transactions on audio, speech, and language processing*, 14(5):1526–1540.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Vũ Minh Quang, Laurent Besacier, and Eric Castelli. 2007. Automatic question detection: prosodic-lexical features and crosslingual experiments. In *Eighth Annual Conference of the International Speech Communication Association*.

Yuexi Ren, Sung-Suk Kim, Mark Hasegawa-Johnson, and Jennifer Cole. 2004. Speaker-independent automatic detection of pitch accent. In *Speech Prosody 2004, International Conference*.

Andrew Rosenberg. 2009. *Automatic detection and classification of prosodic events*. Columbia University.

Magnus Sahlgren. 2008. The distributional hypothesis. *Italian Journal of Disability Studies*, 20:33–53.

Licia Sbattella, Roberto Tedesco, and Alessandro Trivilini. 2014. Forensic examinations: Computational analysis and information extraction. In *International Conference on Forensic Science-Criminalistics Research (FSCR)*, pages 1–10.

Fabio Tamburini and Petra Wagner. 2007. On automatic prominence detection for german. In *Eighth Annual Conference of the International Speech Communication Association*.

Yaodong Tang, Yuchen Huang, Zhiyong Wu, Helen Meng, Mingxing Xu, and Lianhong Cai. 2016. Question detection from acoustic features using recurrent neural network with gated recurrent unit. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 6125–6129. IEEE.

Paul A Taylor. 1993. Automatic recognition of intonation from f0 contours using the rise/fall/connection model.

CW Wightman and Mari Ostendorf. 1991. Automatic recognition of prosodic phrases. In *Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on*, pages 321–324. IEEE.

Jiahong Yuan and Dan Jurafsky. 2005. Detection of questions in chinese conversational speech. In *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, pages 47–52. IEEE.