Speech Processing and Prosody

Denis Jouvet

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
{denis.jouvet}@loria.fr

Abstract. The prosody of the speech signal conveys information over the linguistic content of the message: prosody structures the utterance, and also brings information on speaker's attitude and speaker's emotion. Duration of sounds, energy and fundamental frequency are the prosodic features. However their automatic computation and usage are not obvious. Sound duration features are usually extracted from speech recognition results or from a force speech-text alignment. Although the resulting segmentation is usually acceptable on clean native speech data, performance degrades on noisy or not non-native speech. Many algorithms have been developed for computing the fundamental frequency, they lead to rather good performance on clean speech, but again, performance degrades in noisy conditions. However, in some applications, as for example in computer assisted language learning, the relevance of the prosodic features is critical; indeed, the quality of the diagnostic on the learner's pronunciation will heavily depend on the precision and reliability of the estimated prosodic parameters. The paper considers the computation of prosodic features, shows the limitations of automatic approaches, and discusses the problem of computing confidence measures on such features. Then the paper discusses the role of prosodic features and how they can be handled for automatic processing in some tasks such as the detection of discourse particles, the characterization of emotions, the classification of sentence modalities, as well as in computer assisted language learning and in expressive speech synthesis.

Keywords: Prosody, Speech processing, Prosodic features, Fundamental frequency

1 Introduction

In speech communication, prosody conveys various types of information over the linguistic content of the messages. For example, prosody structures the utterances, thus playing a role similar to punctuation in written texts; and provides ways to emphasize words or parts of the messages that the speacker think are important. Prosody also conveys information on the speaker's attitude and emotional state.

The prosody of the speech is often neglected in automatic speech recognition as well as in manual transcription of speech corpora. On the other side expressive speech is now attracting more and more interest in some speech sciences, such as for speech synthesis [44] and for automatic recognition of emotions [32]. For a long time text-to-speech (TTS) synthesis research was focused on delivering good quality and intelligible speech. Such systems are currently used in information delivery services, as for example in call center automation, in navigation systems, and in voice assistants. The speech

style was then typically a “reading style”, which resulted from the style of the speech data used to develop TTS systems (reading of a large set of sentences). Although a reading style is acceptable for occasional interactions, TTS systems should benefit from more variability and expressivity in the generated synthetic speech, for example, for lengthy interactions between machines and humans, or for entertainment applications. This is the goal of recent or emerging research on expressive speech synthesis.

Prosody is a suprasegmental information , i.e., is defined on segments larger than the phones. Several variables are used to characterize the prosody. This includes the fundamental frequency, the duration of the sounds, and the energy of the sounds. Most of the time it is the evolution of these variables over time, or their relative values that bring prosody information.

Forced speech-text alignment is used to obtain word and phone segmentations of speech signals. Assuming that a precise transcription is available, forced speech-text alignment provides good segmentation results on clean speech signals. However there exists conditions where performance degrades, as for example, on noisy signals, or when dealing with dysfluencies of spontaneous speech, or when processing non-native speech. Similarly, many algorithms have been developed for computing the fundamental frequency. They work well on good quality speech signals, but their performance degrades on noisy speech signals.

The paper is organized as follows. Section 2 details the features and their automatic computation. Section 3 deals with the reliability of the prosodic features. Section 4 discusses the use of prosodic features in various speech applications. Finally a conclusion ends the paper.

2 Computing prosodic features

The computation of the prosodic parameters involves the computation of the phone duration, of the fundamental frequency, and of the phone energy.

2.1 Phone duration

The phone duration is determined from a phonetic segmentation of the speech signal. Such segmentation can be done manually using some speech visualization tool such as Praat [10], or automatically using forced speech-text alignment procedures. Although automatic speech-text alignment provides good results on clean speech data, some manual checking and corrections may be necessary, especially when dealing with spontaneous speech if all speech dysfluencies are not marked in the transcription and properly processed, when processing non-native speech, or in noisy conditions.

2.2 Fundamental frequency

The fundamental frequency (F0) is an important prosody feature. It corresponds to the frequency of vibration of the vocal folds. Many algorithms have been developed in the past to compute the fundamental frequency of speech signals, they are generally referred to as pitch detection algorithms. Several algorithms operate in the time domain. This is

the case of those based on the auto-correlation function (ACF) [9], of the robust algorithm for pitch tracking (RAPT) [52], of the YIN approach [13] and the time domain excitation extraction based on a minimum perturbation operator (TEMPO) [28], [27]. Some algorithms operate in the frequency domain as the sawtooth waveform inspired pitch estimator (SWIPE) [12]. Other algorithms combine processing in the time and in the frequency domains. This is the case of the pitch detection of the Aurora algorithm [49] initially developed for distributed speech recognition, and the nearly defect-free F0 (NDF) estimation algorithm [26]. More recently, new algorithms have also been released, as for example the robust epoch and pitch estimator (REAPER). A pitch tracker has also been developed for automatic speech recognition of tonal languages within the Kaldi toolkit [18]. Their accuracy and reliability is discussed later in section 3.3

2.3 Phone energy

The raw local energy of speech signals is quite easy to compute, and is part of many sets of acoustic features. However getting the phone energy implies some choices: should it be an average value over the whole phone segment, or and estimation in the middle of the phone segment. What is the impact when applied to non stationary sounds such as plosives and diphthongs. Errors on the phone boundaries will also affect the estimation.

Other phenomena must also be taken into account. The energy of the speech signal not only depends on the speaker, but is also dependent on the distance and position between the speaker's mouths and the microphone, on the type of microphone and on the transmission channel. All these variability sources complicates the actual usage of the energy feature. Comparing phone energy between sounds that belong to the same utterance is reasonable, as we can assume that the above acquisition factors do not vary too much within an utterance. However comparing phone energy between speech utterances collected in different conditions may not be reliable, and can lead to unexpected results.

3 Reliabity of prosodic features

This section discusses the reliability of the prosodic features, especially when computed automatically on spontaneous speech, on non-native speech, or on noisy data.

3.1 Speech-text alignments

Speech-text alignment relies on matching the speech signal with a sequence of acoustic models that corresponds to the possible pronunciation variants of the corresponding text. Hence a correct prediction of the pronunciation variants of the words is critical. Usually pronunciation variants are extracted from available pronunciation dictionaries for words present in those dictionaries, and using some grapheme-to-phoneme converters for other words. Well known approaches of grapheme-to-phoneme converters are based on joint multigram models [8], on weighted finite-states transducers [38], on conditional random fields [19], and more recently on neural networks, either long-short

term memory recurrent neural networks [43] or sequence-to-sequence neural net models [56]. In practice it is important to predict all possible pronunciation variants. As any individual grapheme-to-phoneme converter may make mistakes, it is interesting to combine several converters. Predicting the pronunciation variants of names of persons, locations, etc., is more complicated, in particular when dealing with foreign names, which can be pronounced as in the original language, or pronounced using pronunciation rules of the current language, or a mix of both. Some papers have investigated using the origin of the proper names in the prediction process [20].

The speech signal is affected by many variability sources [7], which include speaker, environment noise, channel transmission, etc. Frequently strong accent or non-native accent implies non-standard pronunciation variants, consequently this will introduce mismatches in the alignment process; unless specific pronunciation variants are taken into account. Nevertheless, it should be noted that it is almost impossible to predict all possible non-native pronunciation variants of each word, as non-native variants depends on both the mother tongue and the target language. This would lead to too numerous variants which will be harmful for the alignment process. Spontaneous speech dysfluencies, such as false starts implies matching portions of the signal with partial pronunciation of words that are not always properly predicted. Automatic alignment performance also degrades on noisy signals, which are typical of spontaneous speech signals. Other problems come the manual transcription of the speech signals, which may contain some spelling errors, and some unforeseen notations due to variability in annotation protocols [17].

Most of the speech-text alignment systems relies on acoustic Markov models (with Gaussian mixture models or hybrid approaches with neural network models). In both cases, the structure of the model is a three-state model, which means that are least three acoustic frames must be aligned with each phone model. Consequently this implies a minimum duration of three frames for each phone segment. Conventional acoustic analysis compute frames every 10 ms, leading to a minimum duration of 30 ms, which appears to be too long in some cases in rapid speaking styles. This lead to investigating the usage of a smaller frame shift (5 ms instead of 10 ms) for speech-text alignment [21]. It should also be noted that parametric speech synthesis systems such as HTS [57] and MERLIN [55] relies also on 5 ms frame shifts.

3.2 Phone duration

The duratino of the phones are obtained from the phone segmentation. Hence the quality of the estimated duration of the phones depends on the accuracy and precision of the phone boundaries. For automatic speech-text alignments, the precision of the boundaries depends on the frame shift used: either 10 ms or 5 ms shift. The accuracy depends on the quality of the acoustic models used. Also, some boundaries are clearly marked in the spectrum space, as for example between vowels and fricatives or plosives. On the opposite, boundaries between vowels and sem-vowels or liquids are much less obvious, and often their position is error prone in automatic alignments.

In many cases the automatic speech-text alignment relies on a two step process. A first alignment is carried out using context dependent phone models. Such models provide a refine modeling of the contextual influence of adjacent phones, and thus are

relevant to find the best pronunciation variant for each word occurrence. Once the best pronunciation variant has bee determined for each word occurrence a new alignment is carried out using context-independent phone models, as such models lead to a better determination of time position of the phone boundaries.

3.3 Fundamental frequency

In order to better understand the performance of the various pitch detection algorithms, a set of experiments has been conducted to evaluate and compare their performance on clean speech and on noisy speech [23]. Two speech corpora have been used for the evaluations: the pitch-tracking database from Graz University of Technology (PTDB-TUG) [41] which contains clean English speech signals from 20 speakers, and the SPEECON [1] corpus which contains Spanish speech signals recorded in various real environments from 60 speakers with close-talk and distant microphones placed at different distances from the speakers. This corpora have been developed for pitch tracking evaluation, and are thus provided with reference pitch values.

On clean speech data, large performance variations are observed across speakers, and the average F0 frame error on the PTDB-TUG data varies between 5% and 8% for the 15 approaches that were considered in [23]. According to a recent bibliometric survey [51] the most frequently used pitch detection algorithms are Praat [9], RAPT [52], STRAIGHT [28] [27], YIN [13], and SWIPE [12]. On clean speech signals the ACF algorithm from Praat, and the RAPT algorithms are the two approaches that provides the best performance (average performance of 5An analysis of the results shows that when the level of noise increases, the performance degrades, and the voicing decision is always the main cause of errors. In many cases, the dominant error is the misclassification of voiced frames as unvoiced. Babble noise is also more harmful than the other types of noise. However all algorithms do not behave the same way with respect to the type of noise and the SNR level.

Currently there is no indication of the reliability of the estimated F0 values provided by the various pitch detection algorithms. Some preliminary work has been carried out in this direction [15], but further studies are still necessary.

4 Prosodic features in automatic speech processing

Following the presentation of the prosodic features in Section 2 and a discussion about the reliability of those features in Section 3, this section presents and comments some usage of prosodic features in automatic speech processing.

4.1 Computer assisted language learning

In the last decades there has been enormous progress in the domain of computer assisted foreign language learning (e.g. [16], [53], [48]). When focusing on the pronunciation, the main problem is the automatic detection of mispronunciations. This is achieved using approaches derived from automatic speech recognition technology. Common approaches computes goodness of pronunciation scores [54] which amounts to computing

log likelihood ratio between a forced alignment corresponding to the expected pronunciation and another alignment over an unconstrained phonetic loop. Other approaches introduce frequent mispronunciation variants in the pronunciation lexicon for directly detecting some mispronunciations, and also getting better phonetic segmentation of non-native pronunciations [24].

Besides the correct pronunciation of the expected phones, another aspect to consider is the lexical stress, especially when such phenomenon is not present in the mother tongue. Reliable estimation of the fundamental frequency and of the phone segments is mandatory if one wants to provide relevant feedback to the learner. For example, segmentation reliability of vowels depends on the nature of the adjacent phones [37]. And to avoid providing wrong and useless feedbacks, one should also consider the case where the learner did not pronounce the expected word or expression [40], [11].

4.2 Structuring speech utterances

As mentioned before, prosody helps structuring the speech utterances, thus playing a role similar to punctuation in written texts. Although it is associated with the syntactic structure, the prosodic structure is a priori independent of it.

For the French language, an automatic detection of the prosodic structure has been proposed, based on a theoretical description of prosodic trees; the framework was first developed for prepared speech [36], was later adapted for the semi-spontaneous speech in [46], and further revisited and applied on various types of speech material [3] including spontaneous speech. The approach is based on the assumption that there is a prosodic structure that organizes hierarchically the prosodic groups. Such structure results from contrasts of melodic slopes observed on stressed syllables. Thus, for French, the vowel duration and F0 movements are measured on word final syllables, and the prosodic structure is build by considering the inversion and amplitude of the melodic slopes.

Later an analysis of links between punctuation marks and automatically detected prosodic structures has been conducted on large speech corpora [4] that were manually transcribed and punctuated. Inserting punctuation symbols is somewhat subjective and may vary with annotators. Nevertheless it was interesting to note that more than 85% of the punctuation symbols match with the end of automatically detected prosodic groups.

4.3 Sentence modality

Several studies have been conducted in the past with respect to the detection of the modality of the sentences, as for example for modeling and detecting the discourse structures [25], for distinguishing statements from questions [30], [35], for enriching automatic transcription outputs [29], and for helping creating summaries of meetings [42].

Experiments have been conducted to evaluate various classifications approaches for identifying questions and statements on French data [39]. It was observed that using linguistic features alone provides better results than when using prosodic features only. With linguistic features there is a small drop in performance when using sequences

of words resulting from automatic speech recognition than when using reference transcriptions. However, when dealing with automatic speech transcription data, combining prosodic and linguistic features slightly improved the classification performance.

4.4 Prosodic correlates of discourse particles

Discourse particles are small words or expressions (such as ”well”, ”so”, ”let's see”) that are frequently used in spoken language; they play an important role to steer the flow of the dialogue or to convey various attitudes of the speaker [50]. When such words or expressions are used as discourse particles, their semantic load differ from its usual lexical meaning. Hence the proper detection of discourse particles is important in some applications, as for example for relevant speech understanding, or for speech translation.

A large set of French speech corpora have been used for investigating some discourse particles and for studying their prosodic correlates. The speech corpora used were forced aligned in the ORFEO project[1]. These corpora exhibit various speaking styles ranging from prepared speech (story telling, and broadcast news) to spontaneous speech (interviews and interactions). A set of words that are frequently used as discourse particles in French have been chosen. This include ”*alors*” (”then”, ”what's up”), ”*bon*” (”well”, ”all right”), ”*donc*” (”thus”, ”therefore”), ”*quoi*” (”what”), etc. About 1000 occurrences per word have been randomly selected and annotated as discourse particle or not. It was interesting to observe that the frequency of usage of these words as discourse particles increases significantly with the spontaneity of speech data [2], [22].

The prosodic correlates of these words have been investigated. This include also the position of the word in its prosodic group, determined automatically as described above in Section 4.2. Experiments have also been carried out on automatic classification of the occurrences as discourse particle or not using their prosodic characteristics [14], [22].

When a word occurrence was used as a discourse particle, its pragmatic function was also annotated. The pragmatic function typically indicates the role of the discourse particle in structuring the speech flow, as for example: introduction, conclusion, interruption, etc. Prosodic correlates have been analyzed with respect to the pragmatic functions of the discourse particles [33] As different discourse particles sharing a same pragmatic function often exhibits a set of similar prosodic patterns, experiments have been conducted to investigate their interchangeability [34] when using only textual information, or when using audio plus textual information.

4.5 Expressive speech

Since a few years research on expressive speech synthesis is attracting more and more attention. A few emotions are considered, typically, anger, joy, surprise, sadness, fear and disgust. The speech material necessary to build emotional speech synthesis systems is obtained by having the speaker uttering predefined set of sentences while acting the various emotions. Such approaches lead to good quality emotional speech synthesis systems.

[1] ORFEO project: `http://www.projet-orfeo.fr/`

Such data, in French, has also been used to investigate the differences in the speech signal among the various emotions. Many features vary with the emotion styles: vowel duration, vowel energy, and fundamental frequency. For example the fundamental frequency is on average much higher than in neutral speech for the anger and joy styles, and lower than in neutral speech for the sadness and disgust styles [6]. The range of variation of the fundamental frequency is also much larger for anger speech, and much smaller for sadness speech. With respect to pronunciaton variants, phoneme changes between neutral and emotional speech have been investigated, and a high percentage of schwa omissions has been observed for disgust, fear and joy [5].

Besides investigating the phonetic and prosodic realization of emotional speech, some research is carried out to ease the development of expressive speech synthesis systems using deep learning approaches, and to avoid a specific recording of emotional data from the speech synthesis speaker. Preliminary experiments have investigated the use of transfert learning [31].

For many years, the general approach for the recognition of emotion in speech signals was based on computing a very large set of features on the speech segment, and then providing this huge vector to a classifier [47]. Now deep learning approaches are also used for speech emotion recognition [45].

5 Conclusion

This paper has summarized some research activities relating to prosody in automatic speech processing. After a presentation of the prosodic features, that is the fundamental frequency, the phone duration and the phone energy, we have detailed their computation and discussed their reliability.

In the second part of the paper, we have presented and discussed some research activities dealing with the use of prosodic features. This includes computer assisted language learning, structuring speech utterances, sentence modality, prosodic correlates of discourse particles, and expressive speech.

Forced speech-text alignment and detection of fundamental frequency works rather well on clean speech transcriptions and clean speech signals. However, their performance degrades when dealing with spontaneous speech or noisy signals, which is typical of every day speech. One critical point that received so far very little attention, and needs to be investigated further is the estimation of confidence measures on the computed features. That is, similarly to automatic speech recognition systems that provides confidence measures associated to the recognized words, it would be very useful to have confidence measures associated to the phone segment boundaries, and to the estimated fundamental frequency values. Such confidence measures would be useful for the usage of the prosodic features. For example, in computer assisted language learning, this would allows to obtain a confidence score on the diagnosis, and thus this would lead to much more relevant feedback to the learners.

References

1. Speecon manually pitch-marked reference database for Spanish, ISLRN : 866-498-919-979-7, ELRA ID: ELRA-S0218, Catalogue ELRA (`http://catalog.elra.info/`)

2. Bartkova, K., Dargnat, M., Jouvet, D., Lee, L.: Annotation of discourse particles in French over a large variety of speech corpora. In: ACor4French - Les corpus annotés du français, TALN'2017 - Traitement Automatique des Langues Naturelles. Orléans, France (Jun 2017), https://hal.inria.fr/hal-01585540
3. Bartkova, K., Jouvet, D.: Automatic Detection of the Prosodic Structures of Speech Utterances. In: Železný, M., Habernal, I., Ronzhin, A. (eds.) SPECOM - 15th International Conference on Speech and Computer - 2013. Lecture Notes in Artificial Intelligence, vol. 8113, pp. 1–8. Springer Verlag, Pilsen, Czech Republic (Sep 2013), https://hal.inria.fr/hal-00834318
4. Bartkova, K., Jouvet, D.: Links between Manual Punctuation Marks and Automatically Detected Prosodic Structures. In: Speech Prosody 2014. Dublin, Ireland (May 2014), https://hal.archives-ouvertes.fr/hal-00998031
5. Bartkova, K., Jouvet, D.: Analysis of prosodic correlates of emotional speech data. In: ExLing 2018 - 9th Tutorial and Research Workshop on Experimental Linguistics. Paris, France (Aug 2018), https://hal.inria.fr/hal-01889932
6. Bartkova, K., Jouvet, D., Delais-Roussarie, E.: Prosodic Parameters and Prosodic Structures of French Emotional Data. In: Speech Prosody 2016. Speech Prosody 2016, Boston, United States (May 2016), https://hal.inria.fr/hal-01293516
7. Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C.: Automatic speech recognition and speech variability: A review. Speech Communication (Nov 2007), https://hal.inria.fr/inria-00616506
8. Bisani, M., Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. Speech communication 50(5), 434–451 (2008)
9. Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proc. of the Institute of Phonetic Sciences. vol. 17, pp. 97–110. Amsterdam (1993)
10. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [computer program](2011). Version 6.0.20
11. Bonneau, A., Fohr, D., Illina, I., Jouvet, D., Mella, O., Mesbahi, L., Orosanu, L.: Gestion d'erreurs pour la fiabilisation des retours automatiques en apprentissage de la prosodie d'une langue seconde. Traitement Automatique des Langues 53(3) (2013), https://hal.inria.fr/hal-00834278
12. Camacho, A., Harris, J.G.: A sawtooth waveform inspired pitch estimator for speech and music. The Journal of the Acoustical Society of America 124(3), 1638–1652 (2008)
13. de Cheveigné, A., Kawahara, H.: YIN, a fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America 111(4), 1917–1930 (2002)
14. Dargnat, M., Bartkova, K., Jouvet, D.: Discourse Particles In French: Prosodic Parameters Extraction and Analysis. In: International Conference on Statistical Language and Speech Processing. Budapest, Hungary (Nov 2015), https://hal.inria.fr/hal-01184197
15. Deng, B., Jouvet, D., Laprie, Y., Steiner, I., Sini, A.: Towards Confidence Measures on Fundamental Frequency Estimations. In: IEEE International Conference on Acoustics, Speech and Signal Processing. New Orleans, United States (Mar 2017), https://hal.inria.fr/hal-01493168
16. Eskenazi, M.: An overview of spoken language technology for education. Speech Communication 51(10), 832–844 (2009)
17. Fohr, D., Mella, O., Jouvet, D.: De l'importance de l'homogénéisation des conventions de transcription pour l'alignement automatique de corpus oraux de parole spontanée. In: 8es Journées Internationales de Linguistique de Corpus (JLC2015). Orléans, France (Sep 2015), https://hal.inria.fr/hal-01183352

18. Ghahremani, P., BabaAli, B., Povey, D., Riedhammer, K., Trmal, J., Khudanpur, S.: A pitch extraction algorithm tuned for automatic speech recognition. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. pp. 2494–2498 (2014)
19. Illina, I., Fohr, D., Jouvet, D.: Multiple Pronunciation Generation using Grapheme-to-Phoneme Conversion based on Conditional Random Fields. In: XIV International Conference "Speech and Computer" (SPECOM'2011). Kazan, Russia (Sep 2011), https://hal.inria.fr/inria-00616325
20. Illina, I., Fohr, D., Jouvet, D.: Génération des prononciations de noms propres à l'aide des champs aéatoires conditionnels. In: JEP-TALN-RECITAL 2012. Grenoble, France (Jun 2012), https://hal.inria.fr/hal-00753381
21. Jouvet, D., Bartkova, K.: Acoustical Frame Rate and Pronunciation Variant Statistics. In: International Conference on Statistical Language and Speech Processing. Budapest, Hungary (Nov 2015), https://hal.inria.fr/hal-01184195
22. Jouvet, D., Bartkova, K., Dargnat, M., Lee, L.: Analysis and Automatic Classification of Some Discourse Particles on a Large Set of French Spoken Corpora. In: SLSP'2017, 5th International Conference on Statistical Language and Speech Processing. Le Mans, France (Oct 2017), https://hal.inria.fr/hal-01585567
23. Jouvet, D., Laprie, Y.: Performance Analysis of Several Pitch Detection Algorithms on Simulated and Real Noisy Speech Data. In: EUSIPCO'2017, 25th European Signal Processing Conference . Kos, Greece (Aug 2017), https://hal.inria.fr/hal-01585554
24. Jouvet, D., Mesbahi, L., Bonneau, A., Fohr, D., Illina, I., Laprie, Y.: Impact of Pronunciation Variant Frequency on Automatic Non-Native Speech Segmentation. In: 5th Language & Technology Conference - LTC'11. pp. 145–148. Poznan, Poland (Nov 2011), https://hal.archives-ouvertes.fr/hal-00639118
25. Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., Van Ess-Dykema, C.: Automatic detection of discourse structure for speech recognition and understanding. In: 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings. pp. 88–95. IEEE (1997)
26. Kawahara, H., de Cheveigné, A., Banno, H., Takahashi, T., Irino, T.: Nearly defect-free f0 trajectory extraction for expressive speech modifications based on straight. In: Interspeech. pp. 537–540 (2005)
27. Kawahara, H., Estill, J., Fujimura, O.: Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In: MAVEBA. pp. 59–64 (2001)
28. Kawahara, H., Katayose, H., De Cheveigné, A., Patterson, R.D.: Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity. In: Eurospeech. pp. 2781–2784 (1999)
29. Kolář, J., Lamel, L.: Development and evaluation of automatic punctuation for french and english speech-to-text. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
30. Král, P., Kleckova, J., Cerisara, C.: Sentence modality recognition in french based on prosody. In: International Conference on Enformatika, Systems Sciences and Engineering-ESSE. vol. 8, pp. 185–188. Citeseer (2005)
31. Kulkarni, A., Vincent, C., Denis, J.: Layer adaptation for transfer of expressivity in speech synthesis. In: Proceedings of LTC'2019, 9th Language and Technology Conference (2019)
32. Lanjewar, R.B., Chaudhari, D.: Speech emotion recognition: a review. International Journal of Innovative Technology and Exploring Engineering (IJITEE) 2, 68–71 (2013)
33. Lee, L., Bartkova, K., Dargnat, M., Jouvet, D.: Prosodic and Pragmatic Values of Discourse Particles in French. In: ExLing 2018 - 9th Tutorial and Research Workshop on Experimental Linguistics. Paris, France (Aug 2018), https://hal.inria.fr/hal-01889925

34. Lee, L., Bartkova, K., Jouvet, D., Dargnat, M., Yvon, K.: Can prosody meet pragmatics? Case of discourse particles in French. In: to appear in Proceedings of ICPhS'2019, International Congress of Phonetic Sciences (2019)
35. Margolis, A., Ostendorf, M.: Question detection in spoken conversations using textual conversations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. pp. 118–124. Association for Computational Linguistics (2011)
36. Martin, P.: Prosodic and rhythmic structures in french. Linguistics 25(5), 925–950 (1987)
37. Mesbahi, L., Jouvet, D., Bonneau, A., Fohr, D., Illina, I., Laprie, Y.: Reliability of non-native speech automatic segmentation for prosodic feedback. In: Workshop on Speech and Language Technology in Education - SLaTE 2011. ISCA, Venise, Italy (Aug 2011), https://hal.inria.fr/inria-00614930
38. Novak, J.R., Minematsu, N., Hirose, K.: Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the wfst framework. Natural Language Engineering 22(6), 907–938 (2016)
39. Orosanu, L., Jouvet, D.: Combining lexical and prosodic features for automatic detection of sentence modality in French. In: International Conference on Statistical Language and Speech Processing. Budapest, Hungary (Nov 2015), https://hal.inria.fr/hal-01184196
40. Orosanu, L., Jouvet, D., Fohr, D., Illina, I., Bonneau, A.: Combining criteria for the detection of incorrect entries of non-native speech in the context of foreign language learning. In: SLT 2012 - 4th IEEE Workshop on Spoken Language Technology. Miami, United States (Dec 2012), https://hal.inria.fr/hal-00753458
41. Pirker, G., Wohlmayr, M., Petrik, S., Pernkopf, F.: A pitch tracking corpus with evaluation on multipitch tracking scenario. In: Interspeech. pp. 1509–1512 (2011)
42. Quang, V.M., Castelli, E., Yên, P.N.: A decision tree-based method for speech processing: question sentence detection. In: International Conference on Fuzzy Systems and Knowledge Discovery. pp. 1205–1212. Springer (2006)
43. Rao, K., Peng, F., Sak, H., Beaufays, F.: Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4225–4229. IEEE (2015)
44. Schröder, M.: Expressive speech synthesis: Past, present, and possible futures. In: Affective information processing, pp. 111–126. Springer (2009)
45. Schuller, B.W.: Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM 61(5), 90–99 (2018)
46. Segal, N., Bartkova, K.: Prosodic structure representation for boundary detection in spontaneous french. In: Proceedings of ICPhS. pp. 1197–1200 (2007)
47. Sethu, V., Epps, J., Ambikairajah, E.: Speech based emotion recognition. In: Speech and Audio Processing for Coding, Enhancement and Recognition, pp. 197–228. Springer (2015)
48. Shadiev, R., Hwang, W.Y., Huang, Y.M.: Review of research on mobile language learning in authentic environments. Computer Assisted Language Learning 30(3-4), 284–303 (2017)
49. Sorin, A., Ramabadran, T., Chazan, D., Hoory, R., McLaughlin, M., Pearce, D., Wang, F.C., Zhang, Y.: The ETSI extended distributed speech recognition (DSR) standards: client side processing and tonal language recognition evaluation. In: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. vol. I, pp. 129–132 (2004)
50. Stede, M., Schmitz, B.: Discourse particles and discourse functions. Machine translation 15(1-2), 125–147 (2000)
51. Strömbergsson, S.: Today's most frequently used f 0 estimation methods, and their accuracy in estimating male and female pitch in clean speech. Interspeech 2016 pp. 525–529 (2016)
52. Talkin, D.: A robust algorithm for pitch tracking (RAPT). In: Kleijn, W.B., Paliwal, K.K. (eds.) Speech Coding and Synthesis, pp. 495–518. Elsevier (1995)

53. Viberg, O., Grönlund, Å.: Mobile assisted language learning: A literature review. In: 11th World Conference on Mobile and Contextual Learning (2012)
54. Witt, S.M., Young, S.J.: Phone-level pronunciation scoring and assessment for interactive language learning. Speech communication 30(2-3), 95–108 (2000)
55. Wu, Z., Watts, O., King, S.: Merlin: An open source neural network speech synthesis system. In: SSW. pp. 202–207 (2016)
56. Yao, K., Zweig, G.: Sequence-to-sequence neural net models for grapheme-to-phoneme conversion. arXiv preprint arXiv:1506.00196 (2015)
57. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K.: The hmm-based speech synthesis system (hts) version 2.0. In: SSW. pp. 294–299. Citeseer (2007)