Compilation of Corpora for the Study of the Information Structure–Prosody Interface

Mónica Domínguez[1], Alicia Burga[1], Mireia Farrús[1], Leo Wanner[2,1]
University Pompeu Fabra[1], Catalan Institute for Research and Advanced Studies (ICREA)[2]
C. Roc Boronat, 138
Barcelona, Spain
{monica.dominguez, alicia.burga, mireia.farrus, leo.wanner}@upf.edu

Abstract

Theoretical studies on the Information Structure–prosody interface argue that the content packaged in terms of theme and rheme correlates with the intonation of the corresponding sentence. However, there are few empirical studies that support this argument and even fewer resources that promote reproducibility and scalability of experiments. In this paper, we introduce a methodology for the compilation of annotated corpora to study the correspondence between Information Structure and prosody. The application of this methodology is exemplified on a corpus of read speech in English annotated with hierarchical thematicity and automatically extracted prosodic parameters.

Keywords: prosody, information structure, theme, rheme, thematicity, language resource, annotated corpora

1. Introduction

The interest in the Information Structure–prosody correspondence applied to natural language speech generation lies in the derivation of prosody that is communicatively oriented and, therefore, more expressive. Knowing the linguistic mechanisms involved in human communication is pertinent to the achievement of multifaceted speech technologies that can carry out more complex tasks linked to conversational settings. The so-called Information Structure–prosody interface stands out as a solid ground for starting to build up such a communicative model in the computational field. However, empirical approaches to the Information Structure–prosody interface are scarce, studies on a corpus of more than two speakers are uncommon and the availability of corpora is, so to say, exceptional.

It is usually one aspect of Information Structure that is studied: thematicity. Thematicity defines how content is packaged in terms of "what is being talked about", i.e., the 'theme' and "what is being said", i.e., the 'rheme'. Most of the approaches draw upon a binary flat thematic division and established a one-to-one correspondence between theme–rheme and rising–falling intonation patterns respectively; see, e.g., (Steedman, 2000; Haji-Abdolhosseini and Müller, 2003; Büring, 2003).

A different view on thematicity is that advocated by I. Mel'čuk in the context of the MTT (Mel'čuk, 2001). Compared to the traditional theme–rheme dichotomy, thematicity in the MTT introduces two key features that enhance the scope of the theme–rheme span division, namely: (i) the notion of *specifier*, which sets up the context of the sentence, and (ii) the fact that thematicity is defined over propositions, rather than over sentences. This second feature implies that thematicity is *per se* hierarchical: if a proposition is embedded, its thematicity will be embedded as well. Previous studies proved that hierarchical thematicity corresponds to a wider range of intonation patterns, and is, therefore, a more adequate representation than binary approaches, especially for long syntactically complex sentences; see, e.g., (Domínguez et al., 2016a).

In this paper we present a methodology for the compilation of a corpus for research on the Information Structure–prosody interface from an empirical perspective. This methodology is based on the formal description of information (or communicative) structure by Mel'čuk (2001), which has been already used for the annotation of hierarchical thematicity of written text in (Bohnet et al., 2013); and an automatic annotation of prosody based on a modular pipeline for extraction of acoustic parameters (Domínguez et al., 2016c). An example application is introduced and demonstrated in the online platform *Praat on the Web* (Domínguez et al., 2016b). Classification experiments on a corpus of read speech in English are carried out to validate our approach.

The rest of the paper is structured as follows. The next section presents the motivation and background of this work. The methodology proposed for the compilation of a corpus to study the Information Structure correspondence is described in Section 3. A sample application of a small corpus of read speech in American English is introduced in Section 4. Then, the validation of our approach is presented in Section 5. Finally, conclusions are drawn in Section 6.

2. Motivation and Background

The role of Information Structure (IS) in comprehension of read and spoken speech has been reported for a long time in linguistic and cognitive sciences (Clark and Haviland, 1977; Bock et al., 1983; Fowler and Housum, 1987; van Donselaar and Lentz, 1994). Recent studies in German (Meurers et al., 2011) and Catalan (Vanrell et al., 2013),

for example, show that characteristic intonation patterns that make a distinction between theme and rheme spans contribute to a better understanding of the message.

The relationship between Information Structure and intonation had been discussed even before the **To**nes and **B**reaks **I**ndex (ToBI) (Silverman et al., 1992) was agreed upon as a convention to represent intonation cues. (Beckman and Pierrehumbert, 1986) suggest that the characteristic bitonals for theme and rheme are rising (L*+H) and falling (H+L*), respectively. (Steedman, 2000) proposes a question–answer setting for the identification of theme and rheme and builds upon Beckman's assumption to hypothesize on complete intonation patterns for theme (L*+H LH%) and rheme (H* LL%).

Some attempts have been made on exploring additional aspects of prosody, apart from ToBI contours, in connection with Information Structure. These studies are, as a rule, restricted to one prosodic element in isolation; see, e.g. (Calhoun, 2010) on rhythm (or, rather, 'metrical structure', as the author defines it); (Xu, 1999) on F0 alignment and (Féry, 2013) on prominence and phrasing.

With respect to empirical studies, the intonation of thematicity[1] is studied in German using one speaker (Baumann, 2012). (Féry and Kügler, 2008) study the process of tonal scaling on a corpus of German consisting of eighteen speakers and 2,277 sentences of the same syntactic structure with a varying number of constituents, word order and theme–rheme structure.

All of these studies coincide in that: (i) they analyze only one aspect of prosody, mostly intonation (the variation of F0); and (ii) their representation of binary thematicity is not formalized as required in computational linguistics.

3. Methodology

This paper envisages the compilation of corpora to study the Information Structure–prosody interface from a methodological perspective based on the formal representations of hierarchical thematicity as described by Mel'čuk (2001) and the annotation guidelines established in (Bohnet et al., 2013). The proposed methodology aims to facilitate the following goals:

- to compile large amounts of data from different registers and languages;
- to analyze the hierarchical thematicity–prosody correspondence in human speech using corpus-driven approaches;
- to explore a parametric representation of prosodic elements in its relationship to Information Structure.

Such a methodology addresses two main research issues in this field: (i) the lack of empirical analysis of the IS–prosody correspondence; and (ii) testing of theories on the IS–prosody interface in corpus-based computational models. In order to address these two issues we propose a processing pipeline implemented in the online platform based on Praat (Boersma and Weenink, 2017): *Praat on the Web* (Domínguez et al., 2016b).[2] This platform takes as the basic annotation file, a *TextGrid*,[3] as in standard Praat, and allows scripting of subroutines on both audio and text input using a modular approach, which is not possible in standard Praat.

Figure 1 sketches the pipeline for compilation of corpora to study the IS–prosody interface. The following input is required: (i) a speech (*wav*) file containing the prosodic information ('Pros') and; (ii) the corresponding text (*txt*) file annotated with thematicity ('IS') following the guidelines established in (Bohnet et al., 2013). In module 1, the text is converted to TextGrid format adapting the original annotation to a specific organization into tiers based on the description of thematicity proposed by (Mel'čuk, 2001), as will be detailed in Section 3.1.. Module 2 generates using Praat in-built functions two objects that are needed to extract prosodic information from speech: the pitch and intensity objects. Then, the annotation of prosodic and linguistic features is executed from modules 3 to 6 resulting in an annotated TextGrid with prosodic and thematicity features. Finally, the pipeline outputs a comma separated values (*csv*) file that can be fed to the validation stage to be analyzed by a statistics package or used as input for classification algorithms.

3.1. Annotation of Hierarchical Thematicity

The fact that thematicity, in Mel'čuk's view, is defined over propositions rather than sentences implies that thematicity is *per se* hierarchical, allows embeddedness and, thus, involves different levels of thematicity. For instance, a theme (T1) can be embedded in another theme or rheme (R1) span. Figure 2 shows the levels of embeddedness in example (1), where T1(P2), for instance, is a level 2 theme that belongs to a level 2 proposition (P2) that is embedded in the main T1 span. As more than one thematicity span may exist within the same proposition, abbreviations include a number (e.g., 'SP1') that indicates the number of occurrences at each level (e.g., 'SP2' would be the second specifier in a specific thematicity level).

Guidelines for annotation of hierarchical thematicity were defined and tested in (Bohnet et al., 2013) for the study and annotation of communicative structure in written text. In order to deal with spoken material, an adaptation of these guidelines must be carried out mostly in terms of format to fit in the requirements of the TextGrid format.

[1] A number of other studies refer to thematicity with the term 'givenness'; see, e.g., (Schwarzschild, 1999), and thus talk about 'given' and 'new' information (Chafe and Li, 1976; Clark and Haviland, 1977; Brown, 1983).

[2] `http://kristina.taln.upf.edu/praatweb/`

[3] A dedicated format of Praat for annotation of speech that maps a minimum of one tier to the whole time-stamp of the associated sound file.

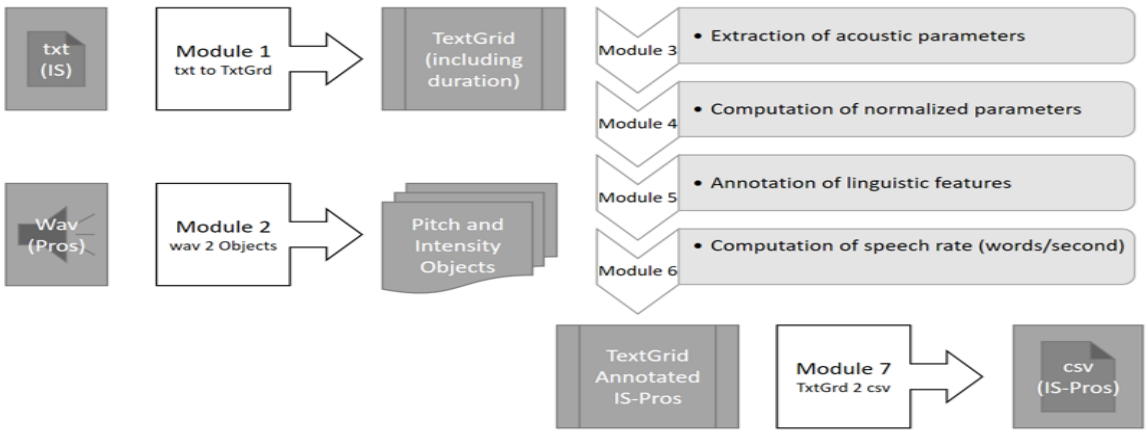

Figure 1: Processing pipeline for IS-Prosody corpus compilation.

(1) [Men {[who]T1(P2) [have played hard all their lives]R1(P2)}P2]T1 [aren't about to change their habits]R1 , [[he]T1(SP1) [says]R1(SP1)]SP1.

Example (1) shows a sentence annotated with thematicity following the guidelines established in (Bohnet et al., 2013). In this annotation, the main proposition (P1) is often not required for annotation, as it is assumed to be marked by the full stop signalling the end of the sentence. However, P1 is required in the TextGrid format, as it is the segment used for further computation of relative prosodic parameters, as will be introduced in the next section. Therefore, the proposed annotation includes two tiers for each level of representation: one for propositions (e.g., L1P) and one for thematicity labels (e.g., L1T).[4] Figure 2 presents the annotation of example (1) using the visualization tool available in Praat on the Web.

Within the processing pipeline, module 5 annotates linguistic features related to the communicative structure of the text, namely, the number of words in each span, the number of embedded spans it contains and the string of labels associated to that interval at the L1P level.

3.1.1. Annotation of Prosody

Automatic extraction and computation of acoustic parameters is carried out using the extension of Praat for feature annotation and the automatic prosody tagger presented in (Domínguez et al., 2016b; Domínguez et al., 2016c).

Table 1 shows the complete list of absolute and relative acoustic parameters (grouped by the three acoustic elements: F0, intensity, and rhythm), and abbreviations (within brackets) used in this paper.

Absolute values are extracted using different predetermined functions available in Praat. Normalized values relative to the whole sample are computed for each segment of analysis, usually a thematicity span (it may be another segment, e.g., a word). Normalized values for mean absolute values of F0, intensity and speech rate are computed using the 'z-score' normalization. Parameters referring to a time point are computed extracting the point of maximum F0 and minimum intensity respectively and calculating the relative time position in the span with a minmax score. Minmax normalization is computed following the equation 1:

$$minmax.t = \frac{x.t - min.t}{max.t - min.t} \quad (1)$$

where:

x.t = point in time where a peak or valley is located within an interval (e.g., word),

min.t = starting point in time of the corresponding interval, and

max.t = ending point in time of the corresponding interval.

Table 1: Prosodic parameters.

Absolute Parameter	Relative Parameter
mean F0 (F0) standard deviation F0 (std.F0) minimum F0 (min.F0)	z-score F0 (z_F0)
maximum F0 (max.F0)	time point of max.F0 (maxF0.t)
mean intensity (int) standard deviation intensity (std.int)	z-score int (z_int)
minimum intensity (min.int)	time point of min.Int (minInt.t)
maximum intensity (max.int)	
duration (dur)	z-score dur (z_dur)
speech rate in words/sec (sr.w)	z-score sr (z_sr)
speech rate in syllables/sec (sr.s)	

[4] The numbers of the levels are correlative indicating the order in the hierarchy: first (1), second (2), third (3), etc.

Figure 2: Example of Hierarchical Thematicity visualization.

In the minmax normalization, the minimum value is the starting time of the interval, which is mapped to 0, and the maximum value is the ending time of the interval, which is mapped to 1. So, the entire range of time points is mapped to the range 0 to 1. This gives us an idea of the relative time location of the peak within a time segment (in this case a word). In other words, the computed minmax score provides information on the location of the F0 peak ('maxF0.t') and intensity valley ('minint.t'). Thus, if an F0 peak is located within the first half of the time span, it will have a score between 0 and 0.5, and if an intensity valley is located within the second half of the span, the score will be between 0.5 and 1.

4. Example of Application

A selection of 109 isolated sentences from the Penn Treebank (Charniak and al., 2000) (section of the Wall Street Journal) was made from the annotated material used in (Bohnet et al., 2013). The corpus contains not only simple sentences, but also coordination, subordination and the combination of both. This varied syntactic composition is related to the representativeness of communicative structure in terms of: number of thematicity levels (up to three in the corpus); position of spans within the sentence and with respect to each other; and continuity or lack of continuity of spans (in particular, rheme spans can be discontinuous). The corpus has an average of fifteen words per sentence with a minimum of three words and a maximum of thirty. This selection of sentences was recorded in a professional studio by twelve native speakers of American English.

There is a balanced six-to-six distribution of male and female speakers. Participants are assigned an anonymous identifier with the format: *speaker (abbreviated as 'spk') – number (a correlative natural number) – gender ('f' for female or 'm' for male)*, resulting in, e.g., 'spk1f'.

Two datasets are created extracting acoustic parameters from different segments (see Table 2). Acoustic data from all twelve speakers is included in the sentence and thematicity span dataset (abbreviated as SSD and TSD, respectively). The main difference between these two datasets is that in SSD the segments are sentences and the classes to be predicted account for the L1 thematicity of each sentence, whereas in TSD the segments are thematicity spans with their corresponding labels assigned to them.

Table 2: Datasets derived from the corpus of read speech.

Acronym	Dataset Name	Speakers	Attributes	Instances	Classes
SSD	Sentence Span Dataset	12	11	1,308	17
TSD	Thematicity Span Dataset	12	14	6,036	31

5. Validation Experiments

Classification experiments are carried out using the Weka 3.8 Workbench (Hall et al., 2009). A bagging classifier with RepTree is using as classifier with a 10-fold cross-validation configuration. We set out to demonstrate the hypothesis that prosodic parameters are related to thematicity labels at the level of two different partitions: the sentence as a whole and each thematicity span. The objective of these experiments is to observe in isolation prosodic parameters and hierarchical thematicity in order to get a closer insight on their relationship. In these experiments, the correspondence of prosody and thematicity is put to test assuming a bidirectional relation between them, but acknowledging that both of them are dependent upon other linguistic phenomena.

5.1. Prediction of Labels within Thematicity Span

The TSD with all thematicity labels is used to perform the prediction of thematicity labels (a total of thirty-one distinct labels) using as attributes acoustic features and number of words in each span. The purpose of the experiment is to observe the correspondence between hierarchical thematicity and acoustic parameters using all speech samples in our corpus. A ZeroR classifier is used as baseline to evaluate and compare the level of improvement

to the bagging classifier. Table 3 shows average precision (P), recall (R) and F-measure (F) results for the bagging classifier (Bag) and baseline (BL).

Table 3: Absolute improvement classification results.

	Precision			Recall			F-Measure		
	BL	Bag	**AbsImp**	BL	Bag	**AbsImp**	BL	Bag	**AbsImp**
TSD	0.05	0.71	**0.66**	0.22	0.71	**0.49**	0.08	0.70	**0.62**
SSD	0.29	0.73	**0.44**	0.54	0.75	**0.21**	0.38	0.74	**0.36**

5.2. Prediction of labels in each sentence

A second experiment is carried out at the sentence level. For each sentence span, acoustic parameters are extracted to predict the thematicity label sequence at L1 using the SSD (a total of seventeen distinct labels are to be predicted). A simple rule classifier (ZeroR), based on a majority vote, is used as baseline. Classification with ZeroR shows a low precision (P=0.29) and F-measure (F=0.38) while recall is 0.54. Then, a bagging classifier is used and results show a considerable increase in all measures with an average absolute improvement over the baseline of P=+0.44 R=+0.19 and F=+0.36. Table 3 reports precision, recall and F-measure results from this classification.

6. Conclusions and Future Work

The contributions of this approach for the compilation of corpora to study the Information Structure–prosody interface are the following: (i) it is automatized to adapt to format requirements; (ii) it assumes a formal representation of thematicity to annotate text instead of the *ad hoc* perspective taken by traditional approaches; and (iii) it proposes the automatic extraction of prosodic cues related to three prosodic elements.

Resources for the automatic conversion of *txt* hierarchical thematicity to TextGrid format and from TextGrid format to *csv* format as well as *arff* files and pitch and intensity objects derived from the audio files used in the example application are made available in the authors' repository.[5] The material from the example application described in this paper (despite its modest size) is the first annotated resource to study the correspondence in English between prosody and hierarchical thematicity as described by Mel'čuk (2001). Moreover, preliminary experiments (Domínguez et al., 2014; Domínguez et al., 2016a) already proved the adequateness of tripartite hierarchical thematicity over traditional binary approaches and its applicability to prosody enrichment in speech synthesis applications (Domínguez et al., 2017).

An automatic approach for the annotation of hierarchical thematicity based on syntactic dependencies is currently being looked into. This advance combined with the present methodology will foster empirically-grounded models to study the Information Structure–prosody interface and to allow the integration of communicatively-oriented approaches to speech technologies.

7. Acknowledgements

This work is part of the KRISTINA project, which has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under the Grant Agreement number H2020-RIA-645012. It has been also partly supported by the Spanish Ministry of Economy and Competitiveness under the María de Maeztu Unit of Excellence Programme (MDM-2015-0502), and the third author is partially funded by the *Ramón y Cajal* program.

[5]This material as well as the code of each module is available under a Creative Commons GNU v.3 License in the following repository: https://github.com/TalnUPF/compilationISpros/

8. Bibliographical References

Baumann, S. (2012). *The Intonation of Givenness. Evidence from German*. Max Niemeyer Verlag, Berlin, Boston.

Beckman, M. E. and Pierrehumbert, J. (1986). Intonational Structure in Japanese and English. *Phonology Yearbook*, 3:255–310.

Bock, J. K., Mazzella, J. R., and Bock, K. (1983). Intonational marking of given and new information: Some consequences for comprehension. *Memory & Cognition*, 11(1):64–76.

Boersma, P. and Weenink, D. (2017). Praat: doing phonetics by computer [Computer program], http://www.praat.org/, version 6.0.14.

Bohnet, B., Burga, A., and Wanner, L. (2013). Towards the annotation of penn treebank with information structure. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1250–1256, Nagoya, Japan.

Brown, G. (1983). Prosodic structure and the given/new distinction. In A. Cutler et al., editors, *Prosody: Models and Measurements*, pages 67–77. Springer, Berlin, Heidelberg.

Büring, D. (2003). On d-trees, beans, and b-accents. *Linguistics & Philosophy*, 26(5):511–545.

Calhoun, S. (2010). The centrality of metrical structure in signalling information structure: A probabilistic perspective. *Language*, 1(86):1–42.

Chafe, W. L. and Li, C. N. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view in subject and topic. *Subject and Topic*, pages 25–55.

Charniak, E. and al., E. (2000). *BLLIP 1987-89 WSJ Corpus Release 1 LDC2000T43*. Linguistic Data Consortium, Philadelphia.

Clark, H. H. and Haviland, S. E. (1977). Comprehension and the given-new contract. *Discourse production and comprehension. Discourse processes: Advances in research and theory*, 1:1–40.

Domínguez, M., Farrús, M., Burga, A., and Wanner, L. (2014). Towards Automatic Extraction of Prosodic Patterns for Speech Synthesis. In *Proceedings of the 7th International Conference on Speech Prosody*, pages 1105–1109, Dublin, Ireland.

Domínguez, M., Farrús, M., Burga, A., and Wanner, L. (2016a). Using hierarchical information structure for prosody prediction in content-to-speech applications. In *Proceedings of the 8th International Conference on Speech Prosody*, pages 1019–1023, Boston, USA.

Domínguez, M., Farrús, M., and Wanner, L. (2016b). An Automatic Prosody Tagger for Spontaneous Speech. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pages 377–387, Osaka, Japan.

Domínguez, M., Latorre, I., Farrús, M., Codina, J., and Wanner, L. (2016c). Praat on the Web: An Upgrade of Praat for Semi-Automatic Speech Annotation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 218–222, Osaka, Japan.

Domínguez, M., Farrús, M., and Wanner, L. (2017). A Thematicity-based Prosody Enrichment Tool for CTS. In *Accepted to show and tell demonstrations at INTERSPEECH17*, Stockholm, Sweden.

Féry, C. and Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics*, 36(4):680–703.

Féry, C. (2013). Focus as prosodic alignment. *Natural Language & Linguistic Theory*, 31(3):683–734.

Fowler, C. A. and Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26(5):489–504.

Haji-Abdolhosseini, M. and Müller, S. (2003). Constraint-Based Approach to Information Structure and Prosody Correspondence. In *Proceedings of the 10th International Conference on Head-Driven Phrase Structure Grammar*, pages 143–162. CSLI Publications.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).

Mel'čuk, I. A. (2001). *Communicative Organization in Natural Language: The semantic-communicative structure of sentences*. Benjamins, Amsterdam, Philadephia.

Meurers, D., Ziai, R., Ott, N., and Kopp, J. (2011). Evaluating Answers to Reading Comprehension Questions in Context: Results for German and the Role of Information Structure. In *Proceedings of the TextInfer 2011 Workshop on Textual Entailment*, TIWTE '11, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Schwarzschild, R. (1999). Givenness, avoidf and other constraints on the placement of accent. *Natural Language Semantics*, 7(1):141–177.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). TOBI: A Standard for Labeling English Prosody. In *2nd International Conference on Spoken Language Processing (ICSLP 92)*, pages 867–870, Banff, Canada, October.

Steedman, M. (2000). Information structure and the syntax-phonology interface. *Linguistic inquiry*, 31(4):649–689, Fall.

van Donselaar, W. and Lentz, J. (1994). The function of sentence accents and given/new information in speech processing: different strategies for normal-hearing and hearing-impaired listeners? *Language and Speech*, 37(4):375–391.

Vanrell, M., Mascaró, I., Torres-Tamarit, F., and Prieto, P. (2013). Intonation as an Encoder of Speaker Certainty: Information and Confirmation Yes-No Questions in Catalan. *Language and Speech*, 56(2):163–190.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics*, 27(1):55–105.