In *Proceedings of International Conference on Spoken Language Processing (September 18-22, Yokohama, Japan),* vol. 4, 1994, pp. 1931–1934.

PITCH-BASED EMPHASIS DETECTION FOR SEGMENTING SPEECH RECORDINGS

Barry Arons

MIT Media Laboratory
20 Ames St., E15-353
Cambridge, MA 02139, USA
e-mail: barons@media.mit.edu

ABSTRACT

This paper describes a technique to automatically locate emphasized segments of a speech recording based on pitch. These salient portions can be used in a variety of applications, but were originally designed to be used in an interactive system that enables high-speed skimming and browsing of speech recordings.

Previous techniques to detect emphasis have used Hidden Markov Models; emphasized regions in close temporal proximity were found to successfully create useful summaries of the recordings. The new research described herein presents a simpler technique to detect salient segments and summarize a recording without using statistical models that require large amounts of training data. The algorithm adapts to the pitch range of a speaker, then automatically selects the regions of highest pitch activity as a measure of emphasis.

INTRODUCTION

Pitch ("fundamental frequency" or "F0") provides information in speech that is important for comprehension and understanding, and can also be exploited for machine-mediated systems. There are many techniques to determine pitch [1] [2], but there have been few attempts to extract high-level information from pitch for use in segmenting speech recordings.

Work in detecting emphasis [3], locating intonational features [4] [5], and finding syntactically significant hesitations based on pause length and pitch [6] have just begun to be applied to speech segmentation and summarization. The SpeechSkimmer system builds upon these ideas and integrates this type of information into an interactive interface for the high speed skimming and browsing of speech recordings

SpeechSkimmer uses simple speech processing techniques to allow a user to hear recorded sounds quickly, and at several levels of detail. SpeechSkimmer exploits properties of spontaneous speech to automatically structure, select, and present salient audio segments in a time-efficient manner. User interaction, through a manual input device, provides continuous real-time control of the speed and detail level of the audio presentation. SpeechSkimmer incorporates time-compressed speech, pause removal, and non-speech audio feedback to reduce the time needed to listen. SpeechSkimmer presents a multi-level structural approach to auditory skimming, and user interface techniques for interacting with recorded speech. The SpeechSkimmer user interface and a pause-based technique for segmenting recordings are detailed in [7] [8].

This research was performed at Interval Research Corporation and the MIT Media Laboratory, and was funded by Interval Research Corporation.

This paper describes a technique for finding emphasized portions in a speech recording. The algorithm adapts to the pitch range of a talker, and then finds segments of high pitch activity as a measure of emphasis and new topic introductions.

RELATED WORK

It is well known in the speech and linguistics communities that there are changes in pitch under various speaking conditions [9] [10]. The introduction of a new topic often corresponds with an increased pitch range. There is a "final lowering," or general declination of pitch, during the production of a sentence. Sub-topics and parenthetical comments are often associated with a compression of pitch range [11]. Such features have been commonly found within and across native speakers of American English.

Much of the literature on prosody and intonation is based on words and phrases, rather than sentences or paragraphs. However, a variety of investigations have shown the relationship between fundamental frequency, sentence structure, and new topic introductions in speech [12] [13] [14].

A speaker may increase their pitch range to highlight the information in a particular phrase, and the pitch range is expanded at the beginning of a new topic [15]. For example, in an investigation of fundamental frequency and discourse structure, it was found that topic changes were associated with large increases in F0 [16].

An experiment on the perception of "spectrally inverted" speech (where semantic information is not available, but prosodic cues are still present) showed that subjects can locate paragraph and sentence boundaries in conversational speech based only on prosodic cues [17]. Sample utterances of topic changes are shown that begin with a much higher F0 than the preceding speech segments.

EMPHASIS DETECTION

Chen and Withgott trained a Hidden Markov Model (HMM, see [18]) on hand-marked data to detect emphasis based on the pitch and energy content of conversations [3]. Emphasized portions in close temporal proximity were found to successfully create summaries of the recordings. This prosodic approach for extracting high-level information from speech signals is promising as it does not require any lexical recognition or understanding. While HMMs are well understood in the speech recognition community, they are computationally complex statistical models that require significant amounts of training data and thus may not be practical for all applications.

While performing some exploratory data analysis on ways to improve on the Chen and Withgott HMM-based approach, it became clear that significant emphasis information was contained solely in the fundamental frequency. Rather than collect-

ing a large amount of training data for an HMM, it appeared possible to create a much simpler emphasis detector by directly looking for patterns in the pitch.

Automatically finding features such as increased pitch range and final lowering in a speech signal is difficult, as pitch data contains similar features at different scales. Ostendorf says "prosody can operate at multiple levels (e.g., word, phrase, sentence, paragraph), making computational modeling of prosody particularly challenging" [19, p. 315].

As part of the research described in this paper, several techniques were investigated to directly find features in a speech signal (e.g., fitting the pitch data to a curve or differencing the endpoints of contiguous segments); however the prosodic features of interest were difficult to find in a general manner.

EMPHASIS DETECTION ALGORITHM

The initial investigation of pitch-based segmentation was made on recordings that were created during an "off-site" workshop. Talkers introduced themselves and presented a 10–15 minute summary of their background and interests. These monologues were recorded with a lavaliere microphone on a digital audio tape recorder (16-bit data sampled at 48 kHz).

A monologue of a male talker was transcribed and manually annotated with paragraph breaks and emphasized regions by a linguist. Several experiments were performed by visually correlating areas of activity in an F0 plot with this hand-marked transcript of the recording. Areas of high pitch variability were strongly correlated with new topic introductions and emphasized portions of the recording as marked in the transcript.

Figure 1 shows the fundamental frequency for 40 seconds of the recorded monologue. There are several clearly identifiable areas of increased pitch activity. Figure 2 is a close-up of the same data. Note that pitch extraction is difficult [1], and that the resulting data may be noisy and contain anomalous points.

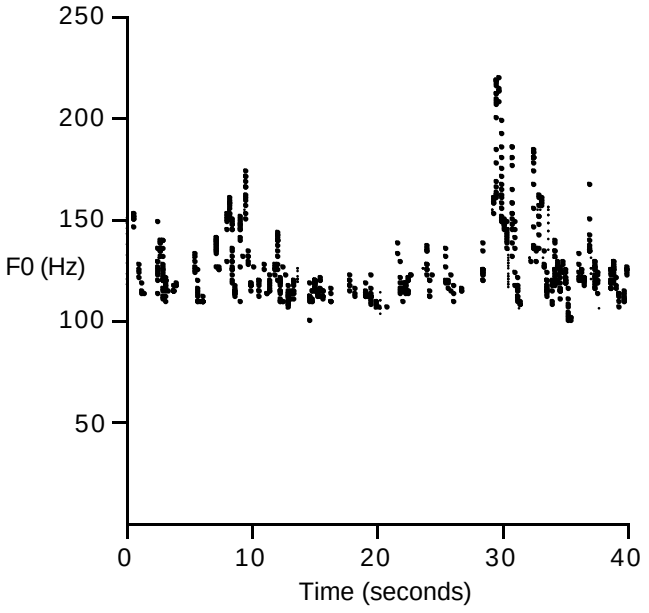

Fig. 1 F0 plot of a monologue from a male talker. Note that the area near 30 seconds appears (and sounds) emphasized. F0 is calculated every 10 ms.

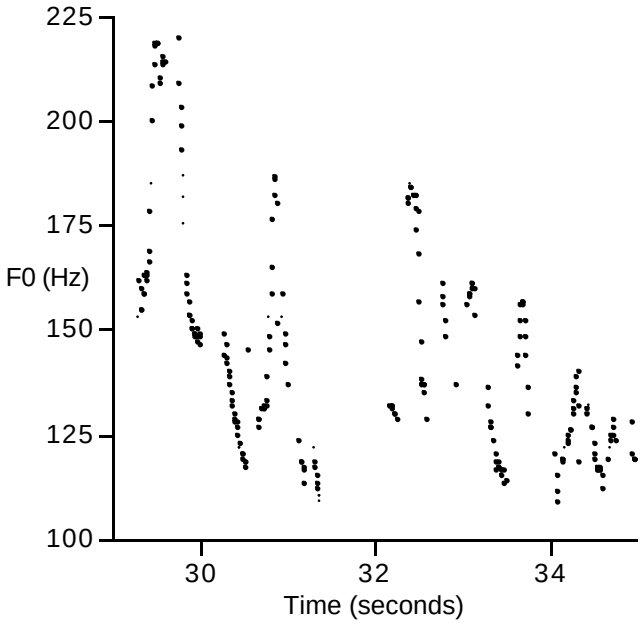

Fig. 2 Close-up of F0 plot in figure 1.

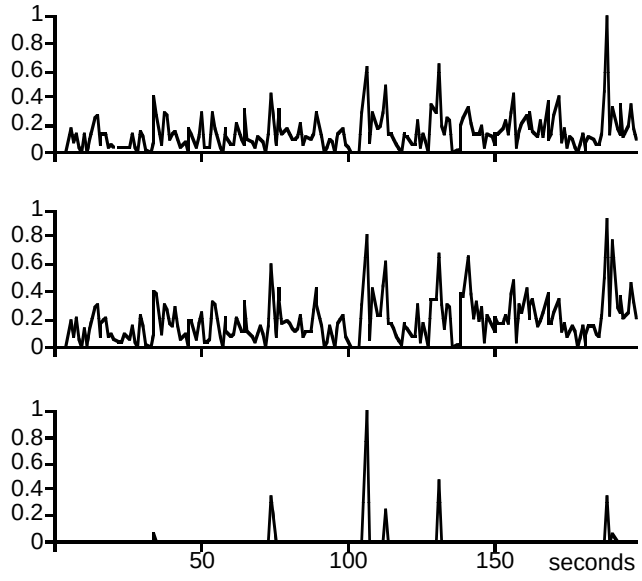

Fig. 3 Comparison of three F0 metrics.

A variety of metrics were generated and manually matched to the hand-marked transcript. The measurements were gathered over one second windows of pitch data (100 frames of 10 ms). One second was chosen to aggregate a reasonable number of pitch values, and to correspond with the length of several words. The F0 metrics evaluated include the mean, standard deviation, minimum, maximum, range, number of frames above a threshold, and number of local peaks within the one second window.

The standard deviation, range, and number of frames above a threshold were most highly correlated with the hand-marked transcript and appeared the most promising for emphasis detection and summarization. Note that these metrics essentially measure the same thing: significant activity and variability in F0. These three metrics thus vary in unison as shown in figure 3. The "number of frames above a threshold" metric was

used in the subsequent development of the algorithm since the resulting data are clean and relatively sparse. However, applying a threshold to the standard deviation or range data provides similar results.

Since the range and baseline pitch vary considerably between talkers, it is necessary to adaptively determine the pitch threshold for a particular talker. A histogram of the pitch data is used to normalize this talker variability. Based on the preliminary investigations, a threshold is chosen to select the pitch frames containing the top 1% of F0 values (figure 4). This threshold was selected as a practical starting point, and can be varied to find a larger or smaller number of emphasized regions. The number of frames in each one second window that are above this threshold are counted to provide a measure of "pitch activity."

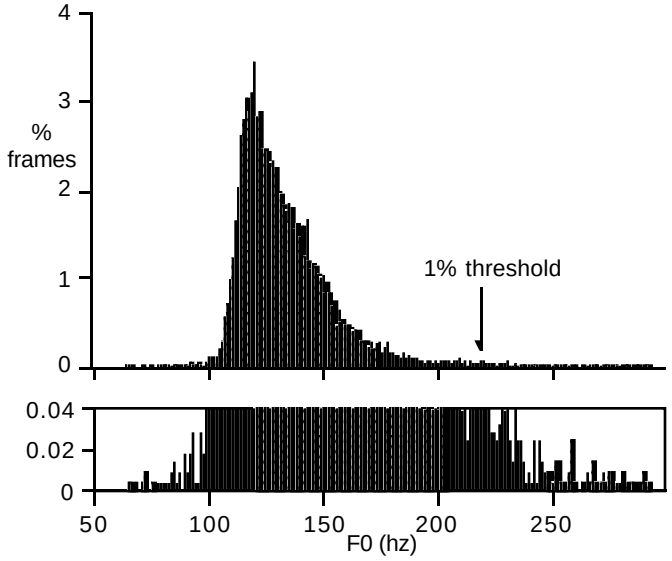

Fig. 4 Pitch histogram for 40 seconds of a monologue from a male talker. The bottom portion has an expanded vertical scale showing the presence of pitch frames above 200 Hz.

The scores of nearby windows are then combined as a measure of emphasis for phrase- or sentence-sized segments of a speech recording. An eight second range was chosen to represent the opening sentences of a new topic introduction. This stage of processing gives extra weight to emphasized areas in close temporal proximity, similar to the method used by Chen and Withgott.

For example, a speech activity of four in window 101 (i.e., four frames above the threshold) would be added to a speech activity of three in frame 106 to indicate there is a pitch activity of seven for the 101–108 second region. This method of combining scores is used instead of collecting speech activity metrics over eight second windows so that the start of the pitch activity can be found at a finer granularity (i.e., one second rather than eight seconds).

It may be desirable to have a set level of compression for a given application. For example, a target of 15:1 compression should select on average one emphasized segment for each two minutes of speech (i.e., 8 seconds out of 120 seconds). If too many speech segments are selected by the algorithm for the desired level of compression, the lowest scoring segments are eliminated.

SAMPLE ANALYSES

Three monologues (one female and two male talkers) were segmented using this pitch-based segmentation technique. The portions selected from the second half of the test recording (only the first half was used in the analysis phase) were highly correlated with topic introductions, emphasized phrases, and paragraph boundaries in the transcript annotated by the linguist.

The four highest scoring segments (i.e., the most pitch activity above the threshold) of each of these recordings were then informally evaluated. People that hear these selected segments generally agree that they are emphasized points or introductions of new topics. The four highest ranking segments (i.e., eight second portions of the original recording) for one of the talkers are:

- OK, so the network that we're building is [pause]… Well this [diagram] is more the VuStation, but the network …
- OK, the second thing I wanted to mention was, the multimedia toolkit. And currently this pretty much something runs on a …
- Currently I'm interested in [pause] computer vision, because I think …
- And then, the third program which is something my group is very interested in and we haven't worked on a lot, is the idea of a news parser …

Along with the stated topic introductions, note the inclusion of the linguistic cue phrases "OK" and "so" that are often associated with new topics [4] [20].

The pitch-based segmentation technique was also applied to a 40 minute lecture and played as part of usability test of the SpeechSkimmer system [8]. The emphasis detection worked well on this recording even though it was of lower quality than the test recording (8 bit linear data sampled at 22.2 kHz).

FUTURE WORK

Some errors are made by the algorithm, such as selecting unimportant portions of a recording, or missing important ones. These errors can probably be reduced by refining the metrics, or combining the emphasis detection with other prosodic information such as pause length.

Along with informal evaluations, such as those described in [8], it is necessary to develop more formalized evaluation methods to extend and refine these speech processing techniques. One of the problems associated with evaluation is in precisely defining the information that one wants to extract from the speech signal. Finding the "major points" in a speech recording is a subjective measure based on high-level semantic and pragmatic information in the mind of the listener. Creating software that can automatically locate acoustic correlates of these features is thus difficult.

Automatically locating "emphasized" or "stressed" portions of a recording is easier (stress can be thought of as an acoustic marker to highlight semantically important words in an utterance [2] [14]), but emphasis is not always correlated with major topics. A talker may use emphasis for a variety of reasons in addition to indicating a new or important point [15]. Some talkers also tend to emphasize just about everything they say, making it hard to identify important segments.

Perhaps the best way to evaluate such a system is to have a large database of appropriately labeled speech data. A variety of hand labeled speech databases are available, but much of the existing labeling has been oriented toward speech recognition

systems rather than high-level information based on the prosody of spontaneous speech (however, see [21]).

Use of such databases allows algorithms to be tested, compared, and improved. For example, it would be possible to evaluate different scoring metrics, such as comparing the use of range or standard deviation against the threshold measure, in an unbiased manner.

Other areas of future work include looking for other pitch-related features in speech. In addition to locating areas of high pitch activity as indicators of new topics, it may be useful to look for features that indicate the end of topics. In the same way that an increased pitch range can indicate a new topic, final lowering can be used to indicate the end of the preceding segment. Pierrehumbert and Hirschberg say "final lowering reflects the degree of 'finality' of an utterance; the more final lowering the more the sense that an utterance 'completes' a topic" [15, p. 279].

CONCLUSION

A technique for automatically locating emphasized portions of a speech signal is described. The algorithm uses simple metrics to measure the pitch activity within a region of a speech recording. The algorithm is straightforward and efficient to implement.

This pitch-based segmentation technique has been successfully used to provide a high-level summary of speech recordings from a variety of male and female talkers. High scoring segments are used in the SpeechSkimmer application to enable efficient skimming of a speech recording. While some errors are made, they are easily navigated around and through using the interactive interface, letting the user find, and listen to, things they are interested in.

Automatically segmenting speech recordings based on acoustic features such as emphasis is important for interactive systems that provide skimming capabilities or attempt to summarize speech recordings. Such techniques are a powerful step toward making it easier and more efficient to listen to recorded speech.

ACKNOWLEDGEMENTS

Meg Withgott helped in my understanding of emphasis detection and supported this research. Lisa Stifelman provided helpful information on prosody and assisted in editing earlier versions of this paper. Thanks also to Michele Covell and Chris Schmandt.

REFERENCES

[1] W. Hess. *Pitch Determination of Speech Signals: Algorithms and Devices*. Berlin and New York: Springer-Verlag, 1983.

[2] D. O'Shaughnessy. *Speech Communication: Human and Machine*. Reading, MA: Addison-Wesley Publishing Company, Inc., 1987.

[3] F. R. Chen and M. Withgott. The Use of Emphasis to Automatically Summarize Spoken Discourse. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing,* IEEE, 1992, pp. 229–233.

[4] J. Hirschberg and D. Litman. Empirical Studies on the Disambiguation of Cue Phrases. *Computational Linguistics 19*, 3 (1993), 501–530.

[5] C. W. Wightman and M. Ostendorf. Automatic Recognition of Intonational Features. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing,* vol. I, IEEE, 1992, pp. I221–I224.

[6] D. O'Shaughnessy. Recognition of Hesitations in Spontaneous Speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing,* vol. I, IEEE, 1992, pp. I521–I524.

[7] B. Arons. SpeechSkimmer: Interactively Skimming Recorded Speech. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST),* ACM SIGGRAPH and ACM SIGCHI, ACM Press, Nov. 1993, pp. 187–196.

[8] B. Arons. Interactively Skimming Recorded Speech. Ph.D. dissertation, MIT, Feb. 1994.

[9] J. Hirschberg and B. Grosz. Intonational Features of Local and Global Discourse. In *Proceedings of the Speech and Natural Language Workshop (Harriman, NY, Feb.23-26),* San Mateo, CA: Morgan Kaufmann Publishers, 1992, pp. 441–446.

[10] K. E. A. Silverman. The Structure and Processing of Fundamental Frequency Contours. Ph.D. dissertation, University of Cambridge, Apr. 1987.

[11] E. J. Kutik, W. E. Cooper, and S. Boyce. Declination of Fundamental Frequency in Speakers' Production of Parenthetical and Main Clauses. *Journal of the Acoustic Society of America 73*, 5 (1983), 17 31–1738.

[12] G. Yule. Speakers' Topics and Major Paratones. *Lingua 52* (1980), 33–47.

[13] W. A. Lea. Prosodic Aids to Speech Recognition. Ch. 8 in *Trends in Speech Recognition*, edited by W. A. Lea. Englewood Cliffs, NJ: Prentice-Hall, 1980.

[14] A. Waibel. *Prosody and Speech Recognition.* San Mateo, CA: Morgan Kaufmann Publishers, 1988.

[15] J. Pierrehumbert and J. Hirschberg. The Meaning of Intonational Contours in the Interpretation of Discourse. Ch. 14 in *Intentions in Communication*, edited by P. R. Cohen, J. Morgan, and M. E. Pollack. Cambridge, MA: MIT Press, 1990.

[16] L. Menn and S. Boyce. Fundamental Frequency and Discourse Structure. *Language and Speech 25*, 4 (1982), 341–383.

[17] J. Kreiman. Perception of Sentence and Paragraph Boundaries in Natural Conversation. *Journal of Phonetics 10* (1982), 163–175.

[18] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE 77*, 2 (Feb. 1989), 257–286.

[19] M. Ostendorf. Prosody (session introduction). In *Proceedings of Human Language Technology (Plainsboro, NJ, Mar 21–24),* San Mateo, CA: Morgan Kaufmann Publishers, 1993, pp. 315–316.

[20] B. J. Grosz and C. L. Sidner. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics 12*, 3 (1986), 175–204.

[21] C. W. Wightman and D. Talkin. Computational Aids for the Study of Prosody (abstract). *Journal of the Acoustic Society of America 95*, 5 Pt. 2 (1994), 2948.