Characteristic Contours of Syllabic-Level Units in Laughter

Jieun Oh[1], *Eunjoon Cho*[2], *Malcolm Slaney*[1,3]

[1]CCRMA, Department of Music, Stanford University, Stanford, USA
[2]Department of Electrical Engineering, Stanford University, Stanford, USA
[3]Microsoft Research, Conversational Systems Research Center, Mountain View, USA
jieun5@ccrma.stanford.edu, ejcho@stanford.edu, malcolm@ieee.org

10.21437/Interspeech.2013-58

Abstract

Trying to automatically detect laughter and other nonlinguistic events in speech raises a fundamental question: Is it appropriate to simply adopt acoustic features that have traditionally been used for analyzing linguistic events? Thus we take a step back and propose syllabic-level features that may show a contrast between laughter and speech in their intensity-, pitch-, and timbral-contours and rhythmic patterns. We motivate and define our features and evaluate their effectiveness in correctly classifying laughter from speech. Inclusion of our features in the baseline feature set for the Social Signals Sub-Challenge of the Computational Paralinguistics Challenge yielded an improvement of 2.4% in Unweighted Average Area Under the Curve (UAAUC). But beyond objective metrics, analyzing laughter at a phonetically meaningful level has allowed us to examine the characteristic contours of laughter and to recognize the importance of the shape of its intensity envelope.

Index Terms: social signals, laughter, acoustic features, classification, Computational Paralinguistics Challenge

1. Background

Despite how essential laughter is in our everyday social interactions, its expressive varieties make it both a rich and challenging subject matter. The social and emotional context surrounding laughter, as well as the characteristic style of the laughing individual, greatly influence the resulting laughter sound [1, 2, 3]. Consequently, there is no concrete laughter archetype or prototype that serves as a starting point against which all laughter instances can be compared for analysis. Granted, we often write out laughter as "haha" in English (or similar forms in different languages), but in reality, laugh sounds rarely consist of repeated syllables that are best transcribed as [ha] or [hɑ].

Another challenge to studying laughter empirically is the difficulty of capturing it "in the wild" in ways that preserve the breadth of its socio-geographical contextual variety. Provine describes the futility of his experience trying to elicit spontaneous laughter in an experiment, as laughter is "a social behavior that virtually disappears in isolated people being scrutinized in a laboratory setting" [1]. Prior studies on laughter typically resorted to funny video clips to induce laughter [4] and used actors to generate laughter sounds to be evaluated by listeners [2, 3]. Such experimental design significantly restricts the types of laughter context that can be studied.

Finally, a lack of common vocabulary has hindered progress on laughter research. Trouvain recognizes this problem and tries to analyze the terminological variety from a phonetic perspective. Trouvain notes that "laughter events are much more complex than implied by an idealized segmentation and most of the existing descriptions of laughter types. More data, clear concepts and more knowledge about the production and acoustics of laughter are necessary to provide phonetically adequate descriptions of the large repertoire of laughter variants" [5].

Amidst such circumstances, the INTERSPEECH 2013 Computational Paralinguistics Challenge – Social Signals Sub-Challenge provides a corpus of spontaneous laughter occurring in 60 phone conversations (involving 120 subjects), along with a set of 141 baseline features and performance results for classification [6]. Specifically, the SSPNet Vocalisation Corpus (SVC) provides 2763 audio clips, each 11 seconds long, that are annotated on every 10ms frame as either laughter, fillers, or garbage; this paper performs classification in the context of this Challenge. Details about the corpus, baseline features, and evaluation criteria are available from the Challenge organizers [6].

2. Approach

Trying to automatically detect laughter in speech raises a fundamental question: Is it appropriate to simply adopt acoustic features – such as the Mel-frequency cepstral coefficients – that have traditionally been used for analyzing linguistic events? Even though laughter shares much of the sound production mechanism as speech, its aural output seems quite musical in expressiveness and variety, displaying a range of movements in pitch, loudness, and timbre [4, 5, 7, 8].

In fact, studies that aim to synthesize laughter using existing speech models have found that they tend to yield unsatisfying results. For instance, Sundaram and Narayanan's two-level model for laughter relies on standard linear-prediction-based analysis–synthesis to generate laughter calls, and naturalness and acceptability evaluation scores of their synthesized clips were significantly below that of real clips [9]. Using a contrasting methodology, Lasarcyk and Trouvain explored imitating conversational laughter with articulatory synthesis and diphone synthesis [10]. They describe limitations of diphone synthesis in emulating breathing and certain laugh syllables that are not available in the predefined phones used for speech; they further note difficulties with using the articulatory synthesis approach in terms of technical limitations (e.g. 1 kPa pulmonic pressure is seemingly not high enough for laughter) and our incomplete knowledge of laughter physiology. If *synthesizing* laughter calls for some kind of specialization or extension to existing speech synthesis models, then we may anticipate an analogous need for *analyzing* laughter.

2.1. Motivation for syllabic-level segmentation

Thus we take a step back and try to characterize the phonetic building-blocks of laughter. But which level of segmentation

is most appropriate for our purpose? Trouvain has proposed phonetic segmentation of laughter at three levels: (1) the segmental level (either consonant or vowel); (2) the syllabic level (the consonant-vowel unit); and (3) the phrasal level, which is a "bout" (a sequence of laughter syllables in one exhalation phase) or an "episode" (the whole laugh) [5].

For the purpose of discriminating laughter from speech, we chose to segment laughter at the *syllabic* level because of the following reasons. First, the syllabic level seems to be high-level enough to be perceptually relevant from the listeners' perspective, but low-level enough to serve as building blocks for modeling an entire laughter episode. Second, segmenting laughter signal at potential syllabic-level boundaries appears to be a straight-forward task, as they are likely to occur at the local minima of signal's energy (by the pulsated exhalation/ inhalation nature of laughter). Finally, our intuition tells us that when humans identify laughter occurring in speech, it is localized at the *boundaries* of syllabic-level units, as opposed to within; thus it seems to work naturally as a meaningful unit for analysis.

2.2. Segmentation method

We segment laughter at every local minima of the signal's energy, which is calculated on 10ms frames and smoothed using a moving average. Specifically, we first calculate the energy e_i for each frame i as a sum of squared sample values in that frame. That is, where F_i is the set of samples in frame i,

$$e(i) = \sum_{n \in F_i} x^2(n) \, .$$

Second, we obtain the smoothed frame-level signal energy $\tilde{e}_i$ (in dB, to better match our perception of loudness) using a simple moving-average rectangular window w of length M (typically between 3-7 frames; determining the optimal M is described in Section 2.3):

$$\tilde{e}(i) = 10 \cdot log \left[\sum_{j=-M/2}^{M/2} w(j) \cdot e(i+j) \right] \, .$$

Finally, we track the delta of smoothed frame energy and segment the signal at every local minima; frames in which the delta of smoothed energy changes from non-increasing to increasing mark the beginning of a new unit S_k. In this manner we segment our signal into approximated syllabic-level units[1].

2.3. Optimal smoothing span (M)

The size of smoothing span, M, directly impacts the resolution of the segmentation. Conceptually, we would like to segment the signal at every syllabic units, following Trouvain's proposal [5]. Without smoothing ($M = 1$), the signal tends to become segmented too finely; if we use too much smoothing (e.g. $M = 10$, which would average 10 frames, or 100ms), then we may miss syllabic boundaries. Based on visual inspection (Figure 1) and classification performance (Figure 2), we determined the optimal smoothing span to be 5 frames.

2.4. Consequence of segmentation on classification

Even though our features (to be described in Section 3) are defined and calculated at the level of segmented units, the Social Signals Sub-Challenge of the Computational Paralinguistics Challenge stipulates that we make classification decisions at a lower level: on 10ms frames. Consequently, *all 10ms frames that belong to the same segmented unit will have identical feature values, and therefore result in the same prediction labels*. While it may seem too crude to ignore all frame-level features, such as those that make up the baseline feature set, our hope is that analyzing at a higher level would offer us a structural view of the signal and capture the notion of a laugh syllable.

[1]Note that scale-space segmentation[11, 12, 13] is a more thorough but computationally demanding alternative to our segmentation method.

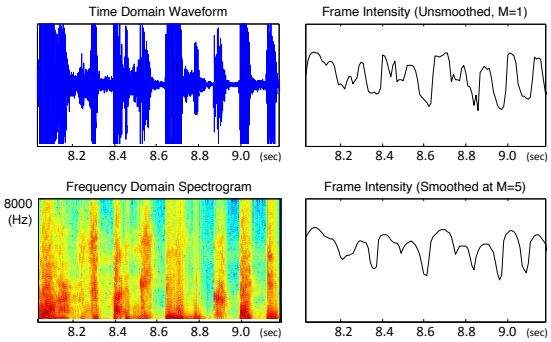

Figure 1: Sample laughter excerpt from SSPNet Vocalisation Corpus: time domain waveform (top left), frequency domain spectrogram (bottom left), unsmoothed frame intensity (top right) and smoothed frame intensity contour (bottom right).

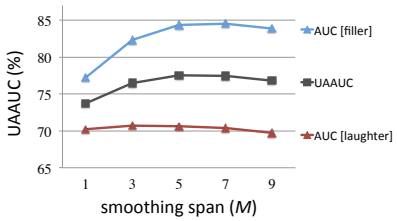

Figure 2: Determining optimal smoothing span as $M = 5$, which yields maximal unweighted average Area Under the Curve (UAAUC) using our syllabic-level feature set

3. Hypotheses

We hypothesize that the syllabic-level units in laughter differ from that of speech in following ways:

1. **Intensity contour**: intensity may be higher and range may be greater [14, 15]
2. **Pitch contour**: a higher mean f0 [4, 7, 15] may result in a different pitch trajectory
3. **Timbral contour**: the spectral envelope may be less variable because laughter predominantly consists of central vowels [4, 16], compared to articulated speech
4. **Rhythmic patterns**: intensity envelope may show rhythmicity around 4-6 pulses per second [4, 7, 8, 17, 18]

4. Syllabic-Level Features

In this section, we motivate and define syllabic-level features that we use to automatically discriminate laughter from fillers and speech. These features allow us to evaluate our hypotheses.

4.1. Intensity contour

The intensity contour focuses on the attack and decay shape of a syllabic unit. Because our analysis units have been segmented at

local minima of signal energy (based on the method described in Section 2.2), the intensity contour of a unit must be arch-shaped: non-decreasing followed by decreasing.

We determine *minimum intensity* (f_1), *maximum intensity* (f_2), and *range intensity* (f_3) based on smoothed intensity values calculated on all 10ms frames that belong to the unit, S_k:

$$f_1 = \min_{j \in S_k} \tilde{e}(j), \quad f_2 = \max_{j \in S_k} \tilde{e}(j), \quad f_3 = f_2 - f_1 \ .$$

Given unit $S_k = \{j_k^1, ..., j_k^N\}$, we define the frame of maximum and minimum intensities as follows:

$$j_k^{\max} = \arg\max_{j \in S_k} \tilde{e}(j), \quad j_k^{\min} = \arg\min_{j \in S_k} \tilde{e}(j) \ .$$

Then we can define *slope to maximum intensity* (f_4) and *slope from maximum intensity* (f_5) to describe the first-order shape of the arch-like intensity contour of a segmented unit:

$$f_4 = \frac{\tilde{e}(j_k^{\max}) - \tilde{e}(j_k^1)}{j_k^{\max} - j_k^1}, \quad f_5 = \frac{\tilde{e}(j_k^N) - \tilde{e}(j_k^{\max})}{j_k^N - j_k^{\max}} \ .$$

In addition, we focus on the intensity delta at the moment of attack and release, as we predicted it to be greater for laughter than for speech as a result of the impulsive, pulsated nature of laughter: *slope at attack* (f_6) is calculated as the delta of intensity between the first two frames of a given unit, and *slope at release* (f_7) is calculated as the delta of intensity between the last two frames of the unit. Note that the shortest possible unit duration is two frames, in which case $f_6 = f_7$.

$$f_6 = \tilde{e}(j_k^2) - \tilde{e}(j_k^1), \quad f_7 = \tilde{e}(j_k^N) - \tilde{e}(j_k^{N-1}) \ .$$

Finally, we calculate syllabic-unit level statistics on the second derivatives of intensity ($\tilde{e}''(j)$): *minimum second-derivative intensity* (f_8), *maximum second-derivative intensity* (f_9), and *mean second-derivative intensity* (f_{10}):

$$f_8 = \min_{j \in S_k} \tilde{e}''(j), \quad f_9 = \max_{j \in S_k} \tilde{e}''(j), \quad f_{10} = \frac{1}{N} \sum_{j \in S_k} \tilde{e}''(j) \ .$$

4.2. Pitch contour

The pitch contour focuses on the melodic line of a laughter syllable. Because the fundamental frequency (f0) is difficult to estimate on frames that are non-harmonic or unvoiced, we recognize that these features may not be robust for certain types of laughter, such as unvoiced inhalations. We nonetheless experiment with describing the pitch contour of syllabic units because certain expressions of laughter are "song-like" [5]. Our features for the pitch contour are calculated using the frame-level f0 values that are taken from the baseline feature set [6], which in turn were obtained using the openSMILE feature extractor [19].

In calculating *minimum f0* (f_{11}) and *maximum f0* (f_{12}), we take the logarithm of the frequency values to better match the perception of pitch in human hearing:

$$f_{11} = \min_{j \in S_k} \log(\text{ f0}(j)\ ), \quad f_{12} = \max_{j \in S_k} \log(\text{ f0}(j)\ ) \ .$$

In order to understand the pitch range of a unit, we calculate *f0 range* (f_{13}). To understand the trajectory of pitch movement, we define *position of minimum f0* (f_{14}) and *position of maximum f0* (f_{15}) as the normalized position (ranging 0.0 to 1.0) of the frame with minimum and maximum f0.

$$f_{13} = f_{12} - f_{11}, \quad f_{14} = \frac{1}{N}\left(\arg\min_{j \in S_k} \text{f0}(j) - j_k^1\right),$$

$$f_{15} = \frac{1}{N}\left(\arg\max_{j \in S_k} \text{f0}(j) - j_k^1\right) \ .$$

4.3. Timbral contour

We take a look at the spectral flux, according to our hypothesis that syllabic-level units in laughter may be characterized by smaller variability in the spectral envelope than speech. We compute *minimum flux* (f_{16}), *maximum flux* (f_{17}), and *mean flux* (f_{18}) as follows, where $s(j)$ is the spectral flux (2-norm between consecutive normalized spectra) calculated on the *smoothed* spectrum[2] of frame j:

$$f_{16} = \min_{j \in S_k} s(j), \quad f_{17} = \max_{j \in S_k} s(j), \quad f_{18} = \frac{1}{N} \sum_{j \in S_k} s(j) \ .$$

4.4. Rhythmic patterns

In addition to the features described above that look at various contours of syllabic units, we included several meta-level features to understand the overall rhythmic and temporal patterns *across* multiple syllabic-level units. These features are meant to capture the specific rhythm that is created by the repeated exhalation and inhalation pulses in laughter (e.g. "ha-ha-ha").

First, we capture the notion of laughter rhythm by calculating the frequency of the signal's intensity contour. We do this by applying a STFT of the intensity contour with a hop size of one frame and window size of W, where W is sufficiently large (e.g. 50 frames) to cover multiple syllabic units. The highest energy at each DFT frame captures the primary modulation frequency of the intensity contour. We average this over the frames in each unit to obtain the *mean intensity-envelope frequency* (f_{19}).

Second, with an *a priori* knowledge that the rhythm of laughter tends to be around 4-6 pulses per second [4, 7, 8, 17], we sum up DFT bins (again, calculated on the signal intensity) that correspond to the 4-6Hz range. The average of these numbers is captured in the *mean laugh-rhythm* feature (f_{20}).

$$f_{19} = \frac{1}{N} \sum_{j \in S_k} \arg\max_m E_j(m) \ ,$$

$$f_{20} = \frac{1}{N} \sum_{j \in S_k} \frac{1}{|F|} \sum_{m \in F} E_j(m) \ ,$$

where F is the set of frequency bins that fall in to the 4-6Hz range, and E_j is the DFT magnitude of a sequence of intensity contours centered at frame j, i.e., $E_j = |\text{DFT}\left[\tilde{e}(j - W/2), ...\tilde{e}(j + W/2)\right]|$.

4.5. Syllabic-level delta and average features

Finally, we calculate delta(Δ) and average(AV) features on each of the syllabic-level features f_1 to f_{20} defined above. Like our features from Section 4.4, these features capture patterns *across multiple syllabic-level units* by tracking changes between two consecutive units (Δ) and averaging feature values across four neighboring units (AV). For the k[th] syllabic unit S_k and a syllabic-level feature f_l, where $1 \leq l \leq 20$:

$$\Delta f_l(S_k) = f_l(S_k) - f_l(S_{k-1}) \ ,$$

$$\text{AV} f_l(S_k) = \frac{1}{5} \sum_{i=-2}^{2} f_l(S_{k+i}) \ .$$

[2] We smoothed the spectrum using a moving average of 20 (out of 256) STFT bins, to roughly follow the formant-peaks.

	DEV SET				TEST SET	
	baseline	syllabic features (w/o Δ & AV)	syllabic features (with Δ & AV)	baseline + syllabic (with Δ & AV)	baseline	baseline + syllabic (with Δ & AV)
	$n = 141$	$n = 15$	$n = 45$	$n = 186$	$n = 141$	$n = 186$
AUC [Laughter]	86.2%	70.9%	74.3%	88.1%	82.9%	85.9%
AUC [Filler]	89.0%	84.3%	87.7%	91.9%	83.6%	84.6%
UAAUC	**87.6%**	**77.6%**	**81.0%**	**90.0%**	**83.3%**	**85.3%**

Table 1: *Results on development set and test set for Social Signals Sub-Challenge, using the same classifier (SVM/SMO) and parameters (c=0.1) as the baselines. Delta(Δ) and average(AV) features track syllabic-level feature values across multiple syllabic-units.*

5. Classification Results

Table 1 summarizes the performance on the Social Signals Sub-Challenge, performing a 3-class (laughter, filler, garbage) classification at every 10ms frames of phone conversations in the SVC corpus. From the 20 syllabic-level features, we removed 5 associated with *f0* because of poor performance. By using our 45 features ($\{f_1$–f_{10}, f_{16}–$f_{20}\}$, and their deltas and averages) on the exactly same classifier (SVM/SMO) and parameters (c=0.1) as the Challenge baselines, we obtained Unweighted Average Area Under the receiver operating Curve (UAAUC) of 81.0% on the development set. Combining our features with 141 frame-level baseline features yielded UAAUC of 90.0% on the development set (exceeding the baseline by 2.4%) and 85.3% on the test set (exceeding the baseline by 2.0%). Because the focus of this work was on investigating syllabic-level features of laughter, we did not experiment with different classifiers or try to tune parameters; however, we recognize that there may be models and parameter settings that would result in a better performance for the given task.

6. Discussion

We analyzed the effectiveness of our syllabic-level features using the WEKA data mining toolkit [20, 21]. Table 2 summarizes our top-ranking features.

rank	feature	type	score
1	f_1	intensity	0.900
2	AV (f_1)	AV intensity	0.868
3	AV (f_{14})	AV intensity	0.843
4	Δ (f_6)	Δ intensity	0.842
5	f_3	intensity	0.798
6	Δ (f_9)	Δ intensity	0.797
7	f_6	intensity	0.790
8	Δ (f_2)	Δ intensity	0.787

Table 2: *Top eight features, using Information Gain Attribute Evaluation in WEKA attribute selection.*

Features describing the intensity-contour of syllabic units were among the top-ranking features, suggesting that there exists a characteristic shape of the syllabic-unit "arch" that can discriminate laughter and fillers from speech. Specifically, Figure 3 (left) illustrates how 'laughter' units (in red) tend to have high *maximum intensity*, and 'fillers' (green) tend to have high *maximum intensity* with extreme (low or high) *minimum intensity*. Having a majority 'garbage' in the lower left corner of the plot is partly an artifact of the nature of the corpus, which labels speech from the speaker on the other side of the phone (and therefore is much quieter) as 'garbage'.

Moreover, Figure 3 (center) illustrates how 'laughter' is unlikely to be at the lowest end of the *mean laugh-rhythm* feature

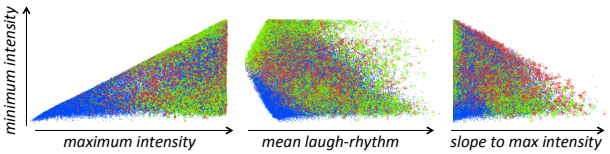

Figure 3: *minimum intensity* versus {*maximum intensity* (left), *mean laugh-rhythm* (center), and *slope to max intensity* (right)} for 'laughter' (red), 'filler' (green), and 'garbage' (blue)

(i.e. left edges), supporting our **hypothesis 4** that laughter exhibits a rhythmic pattern of 4-6 pulses per second. The *Mean laugh-rhythm* feature and its delta are our top-ranking attributes among the non-intensity type features.

Finally, Figure 3 (right) illustrates how 'laughter' units tend to have high *slope to maximum intensity* and high *minimum intensity* (i.e. along the hypotenuse, consistent with **hypothesis 1**). This result intuitively makes sense as laughter syllables tend to be abrupt exhalation pulses with quick rise to intensity peaks.

We note that features for the pitch-contour were our weakest features; this may be a result of having many inharmonic, unvoiced, or otherwise silent frames, causing missing f0 values[3]. Similarly, features describing the timbral-contour (spectral flux) were not as effective as we had hoped, and future work should address how to tease out the timbral characteristics that are uniquely associated with laughter.

7. Conclusion

We have explored how segmenting laughter at a level higher than 10ms frames could aid in characterization of their contours and patterns in ways that improve their automatic detection and classification in a speech context. Furthermore, we observed that incorporating the *changes* (deltas) in syllabic-level features over two consecutive units and *averages* across neighboring units further contribute to the discriminative potential of our feature set. While conceptually analogous to delta features calculated at the frame level (such as many of the features in the baseline set), our syllabic-level delta features track changes in the higher temporal structure of laughter.

Nevertheless, we have so far only computed basic statistics on the intensities, fundamental frequencies, and spectral flux of syllabic-level units. Further research should be conducted to determine more nuanced features that highlight essential acoustic characteristics of laughter. In addition, it would be informative to analyze data using class labels that further distinguish speech laugh from isolated laugh, or exhalation syllables from inhalation syllables, thereby determining features that can discriminate such subclasses of laughter.

[3]Because it is difficult to reliably estimate pitch in casual conversations, a better approach may be to use [22].

8. References

[1] R. R. Provine, *Laughter: A scientific investigation.* Penguin Press, 2001.

[2] D. Szameitat, K. Alter, A. Szameitat, D. Wildgruber, A. Sterr, and C. Darwin, "Acoustic profiles of distinct emotional expressions in laughter," *The Journal of the Acoustical Society of America*, vol. 126, p. 354, 2009.

[3] S. Kori, "Perceptual dimensions of laughter and their acoustic correlates," *Proc. 11th International Congress of Phonetic Sciences*, vol. 4, pp. 255–258, 1989.

[4] J. Bachorowski, M. Smoski, and M. Owren, "The acoustic features of human laughter," *The Journal of the Acoustical Society of America*, vol. 110, p. 1581, 2001.

[5] J. Trouvain, "Segmenting phonetic units in laughter," in *Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona: Universitat Autònoma de Barcelona*, 2003, pp. 2793–2796.

[6] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, marcello Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. Interspeech 2013, ISCA*, Lyon, France, 2013.

[7] K. P. Truong and D. A. Van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, no. 2, pp. 144–158, 2007.

[8] C. Bickley and S. Hunnicutt, "Acoustic analysis of laughter," in *Proc. Int. Conf. Spoken Language Process*, vol. 2, 1992, pp. 927–930.

[9] S. Sundaram and S. Narayanan, "Automatic acoustic synthesis of human-like laughter," *The Journal of the Acoustical Society of America*, vol. 121, p. 527, 2007.

[10] E. Lasarcyk and J. Trouvain, "Imitating conversational laughter with an articulatory speech synthesizer," *Proceedings of the Interdisciplinary Workshop on the Phonetics of Laughter*, 2007.

[11] A. Witkin, "Scale-space filtering: A new approach to multi-scale description," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, vol. 9. IEEE, 1984, pp. 150–153.

[12] R. Lyon, "Speech recognition in scale space," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87.*, vol. 12. IEEE, 1987, pp. 1265–1268.

[13] M. Slaney and D. Ponceleon, "Hierarchical segmentation using latent semantic indexing in scale space," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 3. IEEE, 2001, pp. 1437–1440.

[14] C. E. Williams and K. N. Stevens, "Emotions and speech: Some acoustical correlates," *The Journal of the Acoustical Society of America*, vol. 52, no. 4B, pp. 1238–1250, 1972.

[15] H. Rothgänger, G. Hauser, A. C. Cappellini, and A. Guidotti, "Analysis of laughter and speech sounds in Italian and German students," *Naturwissenschaften*, vol. 85, no. 8, pp. 394–402, 1998.

[16] D. Szameitat, C. Darwin, A. Szameitat, D. Wildgruber, A. Sterr, S. Dietrich, and K. Alter, "Formant characteristics of human laughter," *The Phonetics of Laughter*, p. 9, 2007.

[17] W. Ruch and P. Ekman, "The expressive pattern of laughter," *Emotion, Qualia, and Consciousness*, pp. 426–443, 2001.

[18] L. S. Kennedy and D. P. Ellis, "Laughter detection in meetings," in *NIST ICASSP 2004 Meeting Recognition Workshop, Montreal*. National Institute of Standards and Technology, 2004, pp. 118–121.

[19] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[20] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and techniques*. Morgan Kaufmann, 2005.

[21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[22] M. Slaney, E. Shriberg, and J.-T. Huang, "Pitch-gesture modeling using subband autocorrelation change detection," in *Proc. Interspeech 2013, ISCA*, Lyon, France, 2013.