Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions

John Kane[1], Irena Yanushevskaya[1], Céline de Looze[1], Brian Vaughan[2], Ailbhe Ní Chasaide[1]

[1]Phonetics and Speech Laboratory,
School of Linguistic, Speech and Communication Sciences,
Trinity College Dublin, Ireland
[2]Dublin Institute of Technology, Ireland

10.21437/Interspeech.2014-79

Abstract

For many applications in human-computer interaction, it is desirable to predict between-(gaps) and within-(pauses) speaker silences independently of automatic speech recognition (ASR). In this study, we focus a dataset of 6 dyadic task-based interactions and aim at automatic discrimination of gaps and pauses based on f_0, energy and glottal parameters derived from the speech just preceding the silence. Initial manual annotation reveals strong discriminative power of intonation tune types. In a subsequent automatic analysis using descriptive statistics of parameter contours, as well as a modelling of such contours using principal component analysis, we are able to speaker-independently predict pauses and gaps at an accuracy of 70 % compared to a 56 % baseline.

Index Terms: Spoken interaction, turn-taking, speech timing, prosody, glottal source

1. Introduction

A crucial component of spoken interaction is the timing: when we start/stop talking, when we interrupt, offer feedback etc. Long between-speaker silences (gaps henceforth) can lead to *lapses* or awkward silences, whereas too short gaps can give the impression of a lack of consideration [1]. Similarly, inappropriate interruption of within-speaker silences (pauses henceforth) can be seen as rude. Consequently, timing effectively determines the flow of spoken interactions. Despite this, many spoken dialogue systems use a fixed duration threshold for the computer to begin speaking after the human interlocutor stops [2]. This inevitably leads to a rather unnatural interaction. The use of fixed duration thresholds is due to the difficulty in modelling the varying timing of utterances. Consistency is often preferred to potentially introducing large errors. Nevertheless researchers have sought to improve this timing in spoken dialogue systems.

In some of the early scientific work on this topic, [3] highlighted the importance of syntactic cues for managing turns in conversation. More recent work in support of this, [4] in particular, emphasised the higher importance of syntax over intonational cues for signalling turn-taking. [5] corroborated the relevance of syntactic cues, but also stressed the importance of intonational patterns. [2] looked to exploit this knowledge in an automatic system to allow a dynamically varying duration threshold. By exploiting semantic, prosodic, timing and speaker characteristic based features they were able to reduce latencies by 24 % compared to a fixed duration threshold setup. Despite the claims in [4], some approaches look to avoid using lexical and syntactic information which crucially depends on automatic speech recognition (ASR), the accuracy of which may degrade depending on the recording condition. In terms of prosody, previous studies on English, Finnish and Swedish speech data [6, 7, 8] have reported findings of gaps being predominantly preceded by falling and rising intonation patterns, whereas pauses were more often preceded by flat pitch patterns. Based on these indications some papers have looked to quantify this using automatically derived acoustic parameters. [9] found significant discriminative power of f_0, intensity and voice quality (jitter and shimmer) related features. [10] described a method for integrating prosodic features into an end-of-utterance detection system which looks to exploit the timing information of silences by making repeated predictions at various stages during the silence. Other previous studies have found phrase final lengthening [11] and a measure of f_0 movement parameter [12] as effective indicators of speaker hold/change in conversation. In the present study, we look to extend prosody-based feature extraction from the speech-chunks immediately preceding silences for the purpose of discriminating pauses and gaps. We interpret prosody in a rather broad sense to include glottal-based measurements as well as f_0 and energy parameters. Besides sampling such parameter contours using descriptive statistics, we also attempt to model the shape of these curves using a method based on principal component analysis (PCA). The extracted features are examined through a mutual information based feature assessment before being used as part of speaker independent classification experiments.

2. Speech data

2.1. Recordings

The speech data used in this study comes from the work in [13] (DIT Emotional Speech Corpus). We use audio data from 6 dyadic interactions (involving 6 females and 6 males). The interactions are based on a shipwreck scenario game where participants were presented with 15 items and were given 10 minutes to rank the items in order of usefulness to their survival. Points were received for correct arrangement of individual items and lost for incorrect arrangement. Recordings were carried out with participants in separate isolation booths using a professional Neumann microphone connected to an Apple Mac-based Digidesign Pro-Tools Mbox2 recording system. The audio signal was digitised at 96 kHz/24 Bit and recorded using Pro-Tools software as two separate audio streams. Audio was then downsampled to 16 kHz/8 Bit.

2.2. Automatic annotation of pauses and gaps

The annotation of pauses and gaps is carried out automatically using a similar approach to that described in [1]. Besides avoiding labour-intensive manual annotation, automatic annotation has the benefit of being objective as well as realistic to a real-world system. For the annotation procedure, binary voice activity detection (VAD) was carried out on both speaker channels. We use the VAD algorithm proposed in [14] which is freely available in the VOICEBOX toolkit (`http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`. A schematic output of the VAD is shown in Fig. 1. Pauses and gaps, as well as overlaps can be easily determined from the binary VAD on both speaker channels. A minimum duration of *allowed* pauses is set to 200 ms to avoid false detections for speech events like plosives. Silences below this duration are bridged in the current analysis. Note that decisions on such threshold settings, though necessary for an automatic system, are known to significantly affect the resulting duration distributions [15]. This automatic procedure resulted in 460 gap and 410 pause silences being annotated. This annotation procedure purposely avoids using terms like 'turns' and 'backchannels', which are notoriously difficult to specify from an automation perspective.

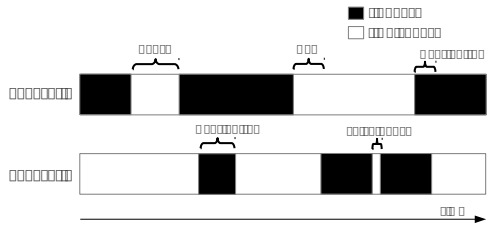
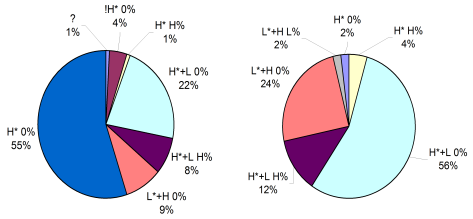


Figure 1: Schematic of dyadic conversation illustrating occurrences of pauses, gaps and overlaps. The 'false pause' indicates a silence which is below the minimum pause duration (here set to 200 ms).

2.3. Human predictable dataset

Due to the nature of how people interact certain pauses and gaps may be *unexpected*. Further, with the application of turn-management of human-computer interaction in mind, it is really those occasions where pauses and gaps are clearly predictable by human listeners that such a dialogue system needs to get right. As a result, in order to derive what we refer to as the *human predictable* subset of the data, we decided to do manual annotation of the automatically extracted dataset. Perceptual analysis of the entire dataset (IPUs preceding the 870 silences detected as pauses or gaps) was carried out by three raters. Each rater was presented with the stimuli in a randomised order and had to indicate, using a 5-point scale, whether a gap (speaker change) or a pause (same speaker continues) follows. The scale was: 1) Very certain the CURRENT speaker continues, 2) Quite certain the CURRENT speaker continues, 3) Don't know! 4) Quite certain the OTHER speaker begins, 5) Very certain the OTHER speaker begins. The raters also had the option to indicate that there was an error in the automatic extraction of stimuli, which may occur due to premature truncation of utterances. In total, under 6 % of stimuli were marked as an error by any of the raters. Inter-rater agreement is quantified using Krippendorf's α, and analysis reveals $\alpha = 0.74$, indicating good agreement across the annotators. To create our *human predictable* dataset, we simply retain samples where all raters agree that it is a pause or a gap (both *very* and *quite* certain included). This resulted in 302 gaps and 288 pauses included in the *human predictable* dataset, which amounts to 70 % of the original dataset (even across pauses/gaps).



(a) Pauses (b) Gaps

Figure 2: Inventories of intonation tune types for the final intonational phrase (IP) in the speech preceding pauses (a) and gaps (b)

2.4. Intonation annotation

To investigate if intonation patterns in the speech preceding pauses and gaps provide any discriminative power, the nuclear tunes (pitch accents and associated boundary tones) of the intonational phrase (IP) closest to the gap/pause silence were annotated using the IViE system [16]. The following tune types were used: H*+L 0% (fall), L*+H 0% (low rise), H*+L H% (fall-rise), L*H L% (rise-fall), H* H% (high rise), !H* 0% (downstep), H* 0% (level).

The annotation was carried out by two raters, who agreed on the labels for 71 % of samples. Given this reasonably high level of agreement, we opt to just use the first rater's annotations in this study. Based on these annotations, Fig. 2 displays the inventory of intonation patterns across the two conditions. The most striking difference is the large proportion of level intonation patterns preceding pauses (55 %) compared to gaps (2 %). Falling (H*L) and rising (L*H) patterns account for more of the utterances before gaps (56 % and 24 %, respectively), however they are not uncommon before pauses (22 % and 9 %, respectively).

3. Feature extraction

3.1. Acoustic parameters

A set of acoustic, prosody-related parameters are extracted from the speech-chunks preceding the pause/gap silences. Various timespans for such analysis have been considered in the literature, typically ranging from 200 ms to 1 s [8, 9], with no real agreement on the optimal timespan for automatic processing. We use 500 ms, as it corresponds roughly to two syllables and may contain sufficient variation in parameter contours to allow prediction of pauses and gaps.

f_0 is extracted using the summation of residual harmonics (SRH) technique described in [17]. f_0 contours are converted to semi-tones (100 Hz reference), and interpolation over unvoiced regions is applied to ensure continuous contours. Signal energy is extracted using 25 ms frames. Two parameters characterising aspects of the glottal excitation are also included. The normalised amplitude quotient (NAQ, [18]) is extracted from glottal source signals, estimated by iterative and adaptive inverse

filtering (IAIF, [19]). The IAIF inverse filtering is done pitch-synchronously, on glottal closure instant (GCI) centred frames of duration twice that of the local glottal period. GCIs are detected using the SE-VQ algorithm [20]. Finally, the recently proposed maxima dispersion quotient (MDQ) parameter [21] is extracted from the linear prediction (LP) residual signal, by characterising the abruptness of the discontinuity at the GCI. All parameter contours are sampled every 10 ms. Note that these parameterisation methods used here are freely available in the COVAREP repository (`http://covarep.github.io/covarep/`). All parameter contours are converted to z-scores, to normalise between-channel differences.

As a baseline feature set we extract 13 Mel-frequency cepstral coefficients (MFCCs) on 32 ms Hann-windowed frames with a 10 ms shift. We also include a measure of *spectral centre of gravity (COG)*, which is measured using: spectral COG $= \frac{\sum_{f=0}^{Fs/2} |X(f)|^2 \cdot f}{\sum_{f=0}^{Fs/2} |X(f)|^2}$ where $X(f)$ is the FFT of Hamming windowed 32 ms speech frame, f is frequency in Hz and Fs is the sampling frequency (16 kHz in the present study). We also include a measure of *spectral similarity*: spectral similarity $=$ cor$\{20log_{10}(|X_n|), 20log_{10}(|X_{n-1}|)\}$ where X_n is the FFT of the n^{th} speech frame (same windowing procedure as above), and cor$\{.\}$ is a function to measure the correlation between two one-dimensional vectors. The spectral similarity measure will remain close to 1 when adjacent spectra display a low level of variation (i.e. when there is little vocal tract and f_0 movement) and will move closer to zero (or minus values) when there is a large variation between frames. We anticipate that *spectral similarity* be useful for characterising the stable vocal tract setting in filled pauses which often precede pause silences.

3.2. Parameter processing

For f_0, energy, NAQ and MDQ, four statistics are derived from each parameter contour: median, standard deviation (std), interquartile range (IQR) and slope. Slope is derived by fitting a first order regression line to the parameter curve. For MFCCs, spectral correlation and COG, just the median value for each coefficient is retained. One shortcoming, however, of just using such descriptive statistics is that many potentially important aspects of the parameter curves (like concavity and convexity) are not characterised. To address this, we applied further analysis of the f_0, energy, NAQ and MDQ contours using an approach resembling functional data analysis. Comparable approaches to this have been applied previously to model variation in intonation contours [22, 23, 24], and similarly to model glottal pulse shapes [25, 26]. The approach involves compiling datasets of parameter contours for pauses and gaps separately. Contours are smoothed by fitting a polynomial to the curve. An iterative procedure, comparable to that used in [22], is used where at each iteration a polynomial is fit to the parameter curve and for each iteration the order of the polynomial is incremented. The iterations are stopped if the residual error falls below a set threshold. Further, the maximum allowed order is set to 6 (i. e. 3 possible inflection points). Each modelled contour then has its mean subtracted. A principal component analysis (PCA) procedure follows. Note that this is a somewhat nuanced application of PCA which is quite different from the standard application for feature dimension reduction. It involves computing the covariance matrix $\mathbf{C}$ of the modelled parameter curve dataset $\mathbf{A}$ using $\mathbf{C} = \frac{1}{M-1} \cdot \mathbf{A}\mathbf{A}^T$ where M is the number of samples in each contour (M = 50, 500 ms sampled every 10 ms). From the covariance matrix $\mathbf{C}$ we then carry out eigendecomposition to extract the eigenvalues λ_i and the corresponding eigenvectors $\mathbf{v}_i$. $\mathbf{v}_i$ are used as the principal components (PCs) and are ranked according to λ_i (descending order). The eigenvalues λ_i are proportional to the variance explained by $\mathbf{v}_i$.

The implementation of this procedure on the dataset used here is done using a leave-one-speaker out approach. Data of a single speaker is held out and the above PCA procedure is applied to the parameter curves derived from the remaining speakers for the pause and gap datasets separately. The three PCs corresponding to the highest three eigenvalues for each dataset are retained, and correspond to the model of the parameter curve. The 6 PCs are then projected onto the held out data (for both pauses and gaps), giving 6 different scores indicating the extent to which each principal component characterises each curve. To illustrate this, the first principal component derived from the f_0 pause dataset is shown in Fig. 3 (top panel) along with its corresponding score plotted as a function of the intonational labels (bottom panel). It can be seen here that this basis function demonstrates a gently rising pattern and can be effective at discriminating various intonational labels. In particular the labels H* H%, H*+L 0% and L*+H 0% are effectively discriminated.

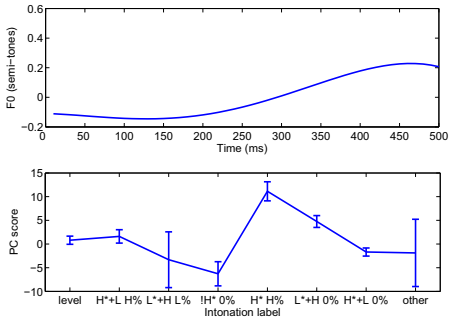


Figure 3: First principal component derived from the pause f_0 set (top panel) and its corresponding score plotted as a function of intonation labels (bottom panel).

3.3. Feature assessment

To investigate the discriminative power of the features included in the present study, we carry out an initial t-test analysis before a mutual information based feature assessment. Multiple independent student t-tests are carried out with each individual feature treated as the dependent variable and pause/gap label as the independent variable. 5 of the 16 descriptive statistics, 10 of the 24 principal component scores and 4 of the 13 MFCCs achieve significance (at $p < 0.05$ level). Also, the spectral similarity feature is found to be highly significant ($p < 0.001$).

For the mutual information analysis, we consider, as in [28], the features, x_i (with i being the feature index), in relation to the discrete class label for pauses and gaps, c. The entropy of classes c can be expressed as: $H(C) = -\sum_c p(c) \log_2 p(c)$, where $p(.)$ is the probability density function (PDF). We model PDFs using a histogram with 30 bins. The mutual information between a given feature, x_i, and the classes, c can be written as: $I(X_i; C) = \sum_{x_i} \sum_c p(x_i, c) \log_2 \frac{p(x_i,c)}{p(x_i)p(c)}$. In the present study we simply consider the *relative intrinsic information* of individual features $\frac{I(X_i;C)}{H(C)}$, which can be interpreted as the proportion of relevant information conveyed by the feature x_i.

Fig. 4 shows the relative intrinsic information of the top 12 features in terms of discriminative power. Spectral similarity achieves the highest score. The rest are mainly f_0 based

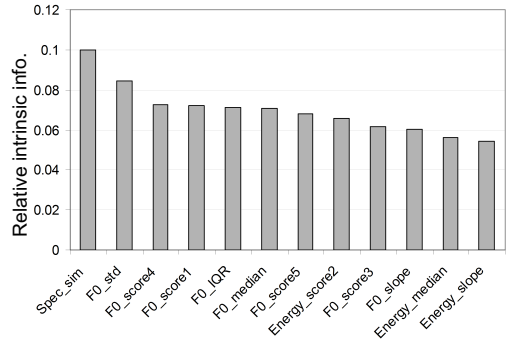


Figure 4: Relative intrinsic information for the top 12 features.

features, including a mixture of descriptive statistics and principal component based measurements. A separate by-speaker assessment (not shown here), reveals a high degree of between-speaker differences, with certain speakers having f_0 based features providing the best discrimination whereas for others it is energy and glottal source based features. Furthermore, the overall amplitude of the relative intrinsic information value varies considerably, with certain speakers having low discriminative power for all features.

4. Classification experiments

4.1. Classifier and experimental setup

We carry out classification experiments to determine the ability of combinations of the features described here to discriminate pauses and gaps. For this we utilise an implementation of Support Vector Machines (SVMs). SVMs in general look to find a separating hyperplane which maximises the functional margin between the two classes. We apply a Radial Basis Function (RBF) kernel [27] which is used to project the feature data into a higher-dimensional space to derive a more effective separating hyperplane. Classification experiments are carried out using a leave-one-speaker-out setup where classifiers are trained on all but one speakers data, and are then tested on the held out data. The held out speaker is then rotated until all speakers have been covered. We experiment with 7 feature sets:

1. **MFCC:** Median values for each of the 13 MFCCs, used as a baseline (13-dimensional)
2. **Descriptive:** Descriptive statistics of f_0, energy and glottal parameter contours (16-dimensional).
3. **PC scores:** Principal component scores of f_0, energy and glottal parameter contours (24-dimensional).
4. **Spectral:** Median values of spectral centre of gravity and spectral similarity (2-dimensional).
5. **All:** All above feature combined (42-dimensional).
6. **Selected:** Top discriminative features (20-dimensional).
7. **Man. intonation:** Manually annotated intonational labels, as a set of binary vectors (8-dimensional).

4.2. Results

The results of the classification experiments are illustrated in Fig. 5. The difficulty of discriminating pauses and gaps is highlighted by the rather high mean classification error (44 %) for the MFCC baseline. The mean error is reduced for descriptive and spectral features (38 %), and even further reduced when using the principal component-based features (35 %). Combining all features does not provide any extra improvement (35 % error), however by selecting the 20 features with the highest relative intrinsic information the mean error reduces to 31 %. A rather high degree of between-speaker variation is observed, with the best performing automatic feature set displaying an accuracy ranging from 80 % to 56 %.

Using the manually annotated intonation labels brings the lowest classification error (18 %), however the accuracy is hugely skewed in the direction of gaps (97 %) compared with pauses (60 %). In fact detection of pauses is achieved at the same accuracy as the best performing automatically derived feature set, however gaps are less accurately detected (75 %).

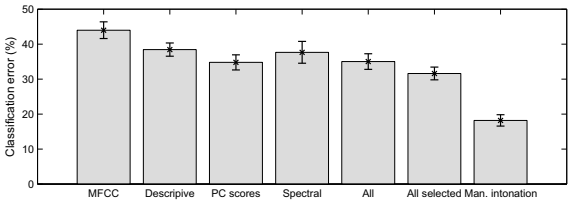


Figure 5: Classification error (%) by feature set.

5. Discussion and conclusion

This paper investigates the potential of prosody-related features in the speech-chunks just preceding silences for discriminating pauses and gaps in task-based dialogues. Findings from preliminary manual analysis of tune types in the intonational phrase just preceding the silence corroborate previous studies [6, 7, 8] with falling and rising intonation patterns mainly preceding gaps, while flat pitch patterns are more characteristic of the speech preceding pauses. The usefulness of a PCA based modelling of parameter contours is demonstrated with regard to separating tune types, and such PCA based features are found to be discriminative of pauses and gaps in both the feature assessment and speaker-independent classification experiments. The spectral similarity feature (proposed as a correlate of filled pauses) is found to be the most discriminative single feature. Low accuracy (56 %) of an MFCC feature set emphasises the difficulty of speaker independent classification of pauses and gaps. A selection of the best features used in the present study provides a considerably improved accuracy of around 70 %. Nevertheless, use of manually obtained intonation tune labels clearly achieves the highest accuracy. This suggests that further attention to automatic labelling of intonation contours may improve detection accuracy. However, one must bear in mind, that manual annotation of pitch patterns was done using the entire inter-pausal unit and not just the final 500 ms, as was used for the automatic feature extraction.

For future work, we intend to utilise this feature set while also exploiting the timing information in silences by constructing a prediction setup similar to that described in [10]. We also intend to further optimise the timespan preceding the silence used for the feature extraction, potentially by automatically identifying pitch targets in the inter-pausal units.

6. Acknowledgments

The first, second, third and fifth authors are supported by the Science Foundation Ireland Grant 09 / IN.1 / I 2631 (FASTNET).

7. References

[1] Heldner, M. and Edlund, J. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38:555–568, 2010.

[2] Raux, A., & Eskenazi, M. Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. *Proceeding of SIGdial, Columbus, OH, USA*, 1–10, 2008.

[3] Sacks, H., Schegloff, E.A., and Jefferson, G., A simplest systematics for the organization of turntaking for conversation. *Language*, 50(4), 696–735, 1974.

[4] de Ruiter, J. P., Mitterer, H., and EnïňĄeld, N. J. Predicting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82, 515–535, 2006.

[5] Ford, C.E., and Thompson, S.A. Interactional Units in Conversation: Syntactic, Intonational, and Pragmatic Resources for the Management of Turns: Chapter in *Interaction and Grammar*, 134–184, 1996.

[6] Local, J. K., Kelly, J., and Wells, W. H. G. Towards a phonology of conversation: turn-taking in Tyneside English. *Journal of Linguistics* 22(2), 411–437, 1986.

[7] Ogden, R. Turn transition, creak and glottal stop in Finnish talk-in interaction. *Journal of the International Phonetic Association* 31(1), 139–152, 2001.

[8] Heldner, M., Edlund, J., Laskowski, K., and Pelce, A. Prosodic features in the vicinity of pauses, gaps and overlaps. *Proceedings of Nordic Prosody*, 95–106, 2009.

[9] Gravano, A. and Hirschberg, J. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25(3):601–634, 2011.

[10] Ferrer, L., Shriberg, E., and Stolcke, A., A prosody-based approach to end-of-utterance detection that does not require speech recognition. *Proceedings of ICASSP*, 608–611, 2003.

[11] Zellers, M. Pitch and lengthening cues to turn transition in Swedish. *Proceedings of Interspeech, Lyon, France*, 248–252, 2013.

[12] Laskowski, K., Edlund, J., and Heldner, M., An instantaneous vector representation of delta pitch for speaker change prediciton in conversational dialogue systems. *Proceedings of ICASSP*, 5041–5044, 2008.

[13] Vaughan, B. *Naturalistic Emotional Speech Corpora with Large Scale Emotional Dimension Ratings*. PhD thesis, Dublin Institute of Technology (DIT), 2011.

[14] Sohn, J., Kim, N. S., and Sung., W. A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3, 1999.

[15] Włodarczak, M., and Wagner, P. Effects of talk-spurt silence boundary thresholds on distribution of gaps and overlaps. *Proceeding of Interspeech, Lyon, France*, 1434–1437, 2013.

[16] Grabe, E., Post, B., and Nolan, F., Modelling intonation variation in English: the IVie system. *Proceedings of Prosody, Poznan, Poland*, 51–58, 2001.

[17] Drugman, T. and Alwan, A. Joint robust voicing detection and pitch estimation based on residual harmonics. *Proceedings of Interspeech, Florence, Italy*, 1973–1976, 2011.

[18] Alku, P., Bäckström, T., and Vilkman, E. Normalized amplitude quotient for parameterization of the glottal flow. *Journal of the Acoustical Society of America*, 112(2):701–710, 2002.

[19] Alku, P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2-3):109–118, 1992.

[20] Kane, J. and Gobl, C. Evaluation of glottal closure instant detection in a range of voice qualities. *Speech Communication*, 55(2):295–314, 2013b.

[21] Kane, J. and Gobl, C. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(6):1170–1179, 2013a.

[22] Su, Z. and Wang, Z. Affective intonation-modeling for mandarin based on PCA. *Computational Linguistics and Chinese Language Processing*, 12(1):33–48, 2007.

[23] Zellers, M., Gubian, M., and Post, B. Redescribing intonational categories with functional data analysis. *Proceedings of Interspeech, Makuhari, Japan*, 1141–1144, 2010.

[24] Arias, J. P., Busso, C., and Yoma, N.B. Energy and f0 contour modeling with functional data analysis for emotional speech detection. *Proceeding of Interspeech, Lyon, France*, 2871–2875, 2013.

[25] Mokhtari, P., Pfitzinger, H. R., and Ishi, C. Principal components of glottal waveforms: towards parameterisation and manipulation of laryngeal voice quality. *Proceedings of VOQUAL, Geneva, Switzerland*, 133–138, 2003.

[26] Chen, G., Kreiman, J., Geratt, B. R., Neubauer, J., Shue, Y-L, and Alwan, A. Development of a glottal area index that integrates glottal gap size and open quotient. *Journal of the Acoustical Society of America*, 133(3) 1656–1666, 2013.

[27] Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, New York, 2006.

[28] Drugman, T., Gurban, M., Thiran, J.-P., Relevant feature selection for audio-visual speech recognition. IEEE International Workshop on Multimedia Signal Processing, 179–182, 2007.