# Speaker Movement Correlates with Prosodic Indicators of Engagement

*Rob Voigt, Robert J. Podesva, Dan Jurafsky*

Linguistics Department, Stanford University, Stanford, CA

robvoigt@stanford.edu, podesva@stanford.edu, jurafsky@stanford.edu

## Abstract

Recent research on multimodal prosody has begun to identify associations between discrete body movements and categorical acoustic prosodic events such as pitch accents and boundaries. We propose to generalize this work to understand more about continuous prosodic phenomena distributed over a phrase - like those indicative of speaker engagement - and how they covary with bodily movements. We introduce *movement amplitude*, a new vision-based metric for estimating continuous body movements over time from video by quantifying frame-to-frame visual changes. Application of this automatic metric to a collection of video monologues demonstrates that speakers move more during phrases in which their pitch and intensity are higher and more variable. These findings offer further evidence for the relationship between acoustic and visual prosody, and suggest a previously unreported quantitative connection between raw bodily movement and speaker engagement.

**Index Terms**: acoustic prosody, visual prosody, movement, gesture, speech-gesture interface, automatic methods

## 1. Introduction

Gesture and movement are fundamental and ubiquitous parts of the human communicative system, but are traditionally understudied phenomena in linguistics. In recent years, interest in the study of multi-modal communication and the connection between speech prosody and "visual prosody" has increased, and empirical evidence has begun to convincingly demonstrate the co-articulatory nature of "gestures and language [as] one system" (McNeill 1992).

For instance, Jannedy and Mendoza-Denton (2006) investigate gesture's role in structuring spoken discourse, showing evidence for the co-occurrence of pitch accents and gestural apices. Krahmer and Swerts (2007) show that even independent of pitch accents, the production of "visual beats" has an effect on the prosodic realization and prominence of the co-occurring speech. Gibbon (2011) demonstrates rhythmic matching between the acoustic and physical beat rhythms in drum-accompanied storytelling in the Ega language. Loehr (2012) confirms findings of temporal synchrony, showing that gestural phrases align with intermediate phrases.

Beyond overt gestures, substantial evidence supports the connection between other kinds of "visual prosody" and speech. Guaïtella et al. (2009) track rapid eyebrow movements in dialogue, and demonstrate their connection with turn-taking in interaction. Cvejic et al. (2010) use facial optical markers in motion capture recordings; in their data, speakers exhibit greater movement during prosodically focused words, even for non-articulatory features such as eyebrow and head movement. Walker (2012) explores "trail-off" conjunctions, showing that speakers and listeners in interaction use visible features such as dropped gaze to signal "disengagement."

Studies in speech perception further demonstrate the important communicative functions of gesture and movement. Munhall et al. (2004) record and recreate 3D models of head movement in talk, finding that subjects are able to correctly identify more syllables when the speech is accompanied by 3D models of natural head movements as compared with distorted or absent models. Scarborough et al. (2009) obtain forced-choice judgments of lexical and phrasal stress from subjects shown video data with the audio track removed; they show that phrasal stress is more easily perceived by subjects than lexical stress. Rilliard et al. (2009) give results for French and Japanese suggesting visual information helps listeners disambiguate "social affects" that are less clear in the audio signal alone.

In spite of this progress, major methodological impediments remain. A principled analysis of gesture in experimental settings requires complex and time-consuming human annotation, commonly based on one of several existing annotation schemes. These include interval annotation of "gesture units" and "gesture phrases" based on written transcripts (Kendon 2004); expressivity annotations on parameters such as "fluidity," "spatial expansion," and "repetitivity" (Chafai et al. 2006); and keyframe-based manual posing of animated 3D characters (Kipp et al. 2007). Though their descriptive power is high, these schemes share the property that they require huge amounts of time and effort from highly-trained human annotators. As a result, existing linguistic studies of gestural prosody necessarily operate on extremely small data sets: for example, Jannedy and Mendoza-Denton (2006) perform their analyses on 130 seconds of videorecorded speech data; Loehr (2012) on speech events from four separate speakers totalling 164 seconds of data.

Analyses of movement pose their own unique difficulties: mainly, raw movement is difficult or impossible to annotate by hand; tools for facial or motion tracking must be used, and this equipment is likely to be expensive or invasive. We therefore know little about the theoretically important relation between affective measures of speaker engagement and the embodied expression that takes form in movement.

This work is aimed at addressing these problems. We introduce a new method for automatically measuring movement magnitude and variance from video data, and apply it to a corpus of single-speaker YouTube videos, extracting acoustic and movement measurements for each phrase in the data. We then investigate the relationship between our proposed movement measure and several prosodic features indicative of engagement including pitch, pitch variance, intensity, and intensity variance (Liscombe, Venditti, and Hirschberg 2003; Mairesse et al. 2007; Gravano et al. 2011; McFarland et al. 2013).

We hypothesize that increased movement amplitude will be predictive of higher values in these acoustic categories. That is, during phrases in which speakers are engaged, excited, and moving more, they will use a correspondingly higher pitch and intensity as well as greater variance in their pitch and intensity.

# 2. Methods

In this work, we propose a pipeline of fully automatic, replicable annotation for the analysis of visual and acoustic prosody on single-speaker videorecorded data.

## 2.1. Data

Web-based video streaming services are an increasingly culturally significant tool for communication. YouTube alone reports more than 100 hours of video uploaded per minute[1], of which a significant portion is certainly linguistic in nature - conversations, vlogs, lectures, and so on. Existing work has begun to use YouTube to investigate multi-modal sentiment (Wöllmer et al. 2013) and sociolinguistic aspects of identity construction (Chun 2013), but this data source remains vastly underutilized.

As we describe in Section 2.4, our new proposed "movement amplitude" measure operates on raw video data, without the need for additional equipment at recording time. This makes it useful for the analysis of movement in YouTube data, which has the significant benefit of replicability: since users who post videos explicitly open them to the public, researchers can apply new methods and test new hypotheses on existing datasets without confronting issues of subject confidentiality.

In this study, we focus on the connection between acoustic and visual prosody in a single, narrowly-defined genre of YouTube videos: the "First Day of School" vlog. In such videos students speak into their cameras to describe their experiences in their first day starting or returning to school. This is a productive genre, with a search for "first day of school vlog" on YouTube returning nearly 1.3 million results.

We collect 14 such videos from different speakers, resulting in a total of 95 minutes of footage. The speakers all fit the most commonly represented demographic in such videos: female high-school-aged students from the United States. We use *pafy*[2] to download 360p-quality mp4-encoded versions of each video. Each video consists of a single speaker seated against a static background. We cut each video to a section of continuous speech, removing introductory and closing title cards.

A first application of our methodology to this dataset is interesting and appropriate for several reasons. The videos were found "in the wild," naturally uploaded by the speaker outside of an experimental context. They are linguistically and gesturally interesting; the speakers are in general very animated and performative, communicating directly to their peer group in discussing social expectations, classes, relationships, and so on.

## 2.2. Automatic Identification of PBUs

We need to extract prosodic units for our analysis; we use pause-bounded units (PBUs), automatically identified with a simple heuristic algorithm using the silence detection function in *Praat* (Boersma and Weenink 2013).[3]

We write a *Praat* script that runs an intensity analysis on the audio track of a given videorecording, then identifies silent and sounding intervals with a minimum duration of one-tenth of a second. We begin with a silence threshold of -30.0 dB and calculate the average length of the phrases thus identified. If the phrases have an average length of greater than two seconds, we increase our silence threshold by +3.0 dB and re-extract, continuing to do so until they average below two seconds in length. Our two-second length threshold is derived from an average length calculation on hand-annotated PBUs from a separate set of interactional data.

## 2.3. Acoustic Features

Following prior work on prosodic engagement, we extract fundamental frequency (henceforth "pitch") and intensity features for each automatically-identified pause-bounded unit in our dataset. We use a *Praat* script to calculate eight acoustic variables for each PBU: a maximum, mean, minimum, and standard deviation for both pitch and intensity.

## 2.4. Movement Amplitude

Here we propose a new measure for the analysis of movement in videorecorded data. The intuition behind this measure is simple. Video data consists of a series of frames, which are fundamentally images, played back at a high speed to simulate movement. In circumstances where the video camera is stationary and the background of the recorded images is relatively static, speaker movement can be quantified by measuring the difference between successive frames.

In practice, we propose a measure obtained by finding the average difference in RGB values between a given pixel and the corresponding pixel in the preceding frame, summing these values across all pixels in the image, and taking the natural log of the total. We extract video frames as uncompressed png images using the ffmpeg software package[4] and compute frame differences using the ImageChops python module from the Python Imaging Library.[5] Formally, we define the movement amplitude (MA) measured at time *t* in (1), with the current frame number *n*, an image size of width *x* and height *y*, and a function $pix_{x,y}(i)$ that returns the red, green, and blue values of a given pixel at an arbitrary frame number *i*:

$$\text{MA}(t) = \ln \sum_{x,y} avg(|pix_{x,y}(n) - pix_{x,y}(n-1)|) \qquad (1)$$

Graphical observation of density and quartile plots of our data confirm the intuition that movement amplitude must be computed in log space. Large movements and gestures are relatively sparse compared with the continuous, generalized movements of speech, and the large variance in the number of pixels they comprise necessitates log space calculation.

Such a measure has many attractive properties. First, it is fully automatic, allowing replicable quantification of visual prosody without painstaking hand-annotation. This means it can be scaled up to provide annotation for datasets of arbitrary size with little additional effort or expense.

Secondly, like measures of acoustic prosody, it is functionally continuous, albeit at a much coarser granularity than most audio recordings. Standard audio sampling rates are 44,100 and 48,000 Hz, while standard video frame rates include 24 and 30 frames per second (FPS). We can calculate one measurement per frame, so these frame rates would allow us to extract movement amplitude samples at 24 and 30 Hz, respectively. In the 30 FPS case this provides one measurement each 33ms. In this study we show that samples extracted at this frequency are sufficient and offer meaningful data on visual prosody; however,

[1] http://www.youtube.com/yt/press/statistics.html

[2] https://pypi.python.org/pypi/pafy

[3] While PBUs provide a meaningful and computationally tractable approximate prosodic unit for this analysis, it remains an interesting task for future work to determine how they might correlate with or be related to a more theoretically principled unit such as the intonational phrase (Pierrehumbert 1980).

[4] http://ffmpeg.org

[5] http://www.pythonware.com/products/pil/

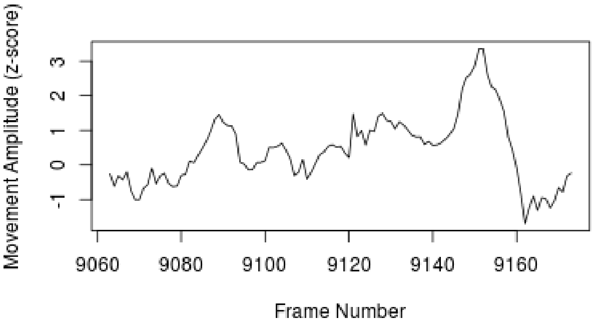

Figure 1: *Visualization of four seconds (120 frames at 30 FPS) of movement.*

the sampling rate of the movement amplitude measure could be arbitrarily scaled in proportion to the quality of the camera available to the experimenter.

Figure 1 demonstrates this continuity, showing movement amplitudes extracted from a four-second clip in our data. The peak near frame 9090, representing an MA measurement one standard deviation above the mean, encodes the combination of a speaker opening her eyes, turning her head towards the camera, and opening her mouth to say "Oh!" in recognition after a moment of thought. The significantly higher peak near frame 9155, three standard deviations above the mean, is primarily the result of a large one-handed swiping gesture.

This plot makes clear another important property of the movement amplitude measure in its current form: it encodes any and all movement, including that of the eyes, mouth, head, body, and so on. Standing up from a seated position, for example, would be recorded as a dramatic peak in MA. It also necessarily compresses 3D movement to a 2D representation in the camera's visual plane. In these ways it is substantially more coarse than any of the measures used in prior studies; however, MA quantifies overall movement in a reasonable way, as the movement of larger objects is given more weight than that of small ones. With computer vision tools such as accurate face detection, in future work this measure could also be applied to submovements to separate out, for example, facial movement as compared to body movement.

Our measure is limited by several required conditions that a recording must meet. The background of the video must be static: the functional result of this is that the contribution to the MA measurement for all pixels that show only background is negligible or absent. Additionally, all speakers in a video must be visually separable. The present study is concerned only with single-speaker video data, but in continuing work we have applied this measure to multi-speaker interactions by defining a rectangular pixel bounding region for each speaker and calculating MA for each speaker only within their region.

These conditions are limiting insofar as other techniques such as hand annotation of gesture could be applied to video footage with variable camera angles and non-static backgrounds. As discussed in Section 2.1, however, data meeting these conditions is reasonable to collect experimentally.

In processing our data we also convert MA measurements to z-scores per speaker to allow for comparable measurements in spite of differences across recording conditions such as speaker distance from the camera, color of the background and speaker clothing, and so on. As with our acoustic variables, for each video we extract an MA maximum, mean, minimum, and standard deviation for each PBU.

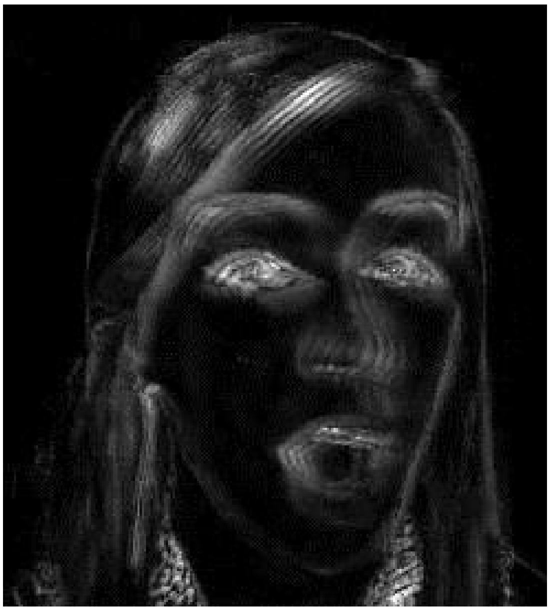

Figure 2: *Five-frame composite visualization of the speaker's head and facial movements as captured by movement amplitude across the first peak shown in Figure 1.*

### 2.5. Statistical Analysis

Automatic extraction of PBUs from our 14 videos results in 2172 observed pause-bounded units, and for each we extract prosodic features for speech and movement as described above.

To model the behavior of this data we use linear mixed-effects models as implemented in the lme4 package in R (Bates et al. 2013). We model speakers as random effects in a series of regressions predicting acoustic variables from our new movement amplitude measure, including log PBU duration in the model as a control variable. The four MA measurements (max, mean, min, and std) are highly collinear, so we use principal component analysis (PCA) for dimensionality reduction. Statistical significance is based on p-values calculated using the lmerTest package in R (Kuznetsova et al. 2013) with degrees of freedom estimated using Satterthwate's approximations.

## 3. Results

Dimensionality reduction with PCA on our MA measurements reveals two orthogonal components that together explain 96% of the variance in the MA data. The loadings for these two factors are seen in Table 1. Factor 1 is interpretable as overall movement, and factor 2 as variance in movement.

| | Overall Movement | Movement Variance |
|---|---|---|
| MA Measure | *Factor 1* | *Factor 2* |
| max | 42.1 | 76.4 |
| min | 73.9 | -54.3 |
| mean | 52.2 | 19.1 |
| std | -6.9 | 29.0 |
| variance explained | 64.8% | 31.2% |

Table 1: *Loadings of orthogonal components for movement amplitude calculated from principal components analysis.*

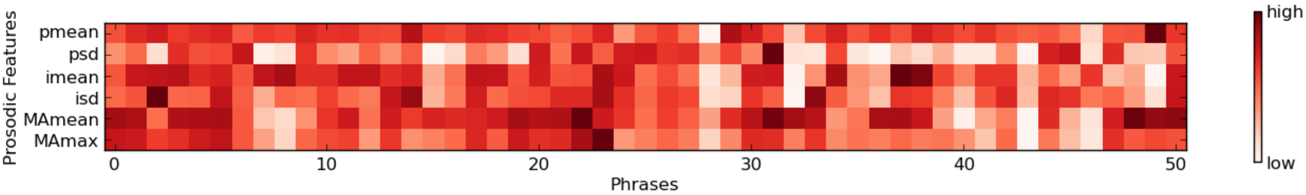

Figure 3: *Visualization of prosodic features including movement across 51 consecutive phrases from one speaker. Each vertical line of boxes represents one spoken phrase, where more deeply-shaded boxes represent higher feature values. Notice light and dark vertical banding; phrases with higher movement amplitude also have higher values for measures of acoustic prosody (and vice versa).*

The results in Table 2 show that high movement amplitude (Factor 1) has a statistically significant positive relationship with all of our acoustic variables except for intensity minimum and pitch minimum. That is, during phrases in which speakers are moving more, we can predict an increase in pitch maximum, mean, and standard deviation as well as intensity maximum, mean, and standard deviation. High movement variance (Factor 2) was not predictive of any of the acoustic features we measured. Though we ran a series of models, the results are highly significant and would survive a Bonferroni correction or any related control for multiple comparisons.

| | OVERALL MOVEMENT | MOVEMENT VARIANCE |
|---|---|---|
| PITCH | *Effect Size (Hz)* | |
| pmax | 4.889 $***$ | — |
| pmin | — | — |
| pmean | 2.797 $***$ | — |
| psd | 0.875 $**$ | — |
| INTENSITY | *Effect Size (dB)* | |
| imax | 0.280 $***$ | — |
| imin | — | — |
| imean | 0.152 $**$ | — |
| isd | 0.082 $***$ | — |

Table 2: *Effect sizes for acoustic features predicted at a statistically significant level by movement amplitude in a series of mixed-effects regressions.*
$*$ *indicates* $p < 0.05$, $**$ *is* $p < 0.01$, *and* $* * *$ *is* $p < 0.001$.
— *indicates no significant relationship.*

To confirm these results, we also run mixed-effects regressions using speaker-scaled pitch max and min, where pitch measurements are converted to a 0-1 scale based on a speaker's overall max and min. We also calculate pitch range features per PBU, which are similarly a value between 0 and 1 calculated by subtracting the scaled pitch max from min. These models show a similar statistically significant trend: an increase of one standard deviation in our overall movement factor predicts use of 1% more of a speaker's pitch range within a given phrase.

## 4. Discussion

Our results build upon prior work to provide further empirical evidence for a strong connection between speech prosody and the "visual prosody" of movement and gesture. In our dataset, phrases with more overall movement were likely to have higher and more variable pitch as well as louder and more variable intensity, confirming our initial hypotheses. This finding is novel in that it adds a dimension of quantity: whereas the previous literature has found primarily temporal synchrony (i.e., the timing of pitch accents aligns with gestural apices), our results demonstrate that more movement is indicative of increased excitement in these prosodic categories.

On the other hand, movement variance - the extent to which a particular phrase had both large movements and periods of little to no movement - was not predictive of any of our acoustic features. This finding is particularly interesting in that a high movement variance would encode some well-defined fully semantic gestures. Consider, for example, a definitive pointing gesture with a pause at its apex. According to the movement amplitude measure, such a gesture would have a high MA value during the stroke and a very low value at the apex, resulting in high MA variation in the PBU during which it occurred. While the movement amplitude measure is not sufficiently fine-grained to make definitive claims in this regard, our findings are suggestive that the synchrony of gestural apices and pitch accents is predominantly temporal and local: that is, a speaker's most "extreme" gestural apices may accurately predict the timing of the immediately adjacent pitch accent, but not necessarily its magnitude with respect to that speaker's global pitch range. This hypothesis remains to be tested in detail in future work.

Additionally, this study makes several valuable methodological contributions to the multi-modal analysis of speech prosody. The newly proposed movement amplitude measure provides a replicable estimation of overall movement from raw video footage, without the need for expensive or invasive equipment at the time of filming or time-consuming annotation effort after the fact. These properties make this measure particularly attractive for the analysis of videos collected "in the wild," such as from internet video sharing sites like YouTube.

The fact that this measure, when combined with automatic extraction of approximate pause-bounded phrases as presented in this paper, is completely automatic for single speakers makes it tractable for empirically-driven sociolinguistic analyses of video data in a way that is simply infeasible by means of hand annotation alone. This paper presented statistically significant results on a narrowly-defined dataset of "First Day of School" vlog posts in order to most directly control for prosodic differences across sociolinguistic groups, but we aim to continue and expand this research. Future work will consider the possible influence of genre effects, social meaning and contextual factors such as performativity, differences in interactional or conversational speech as compared with monologues, and the influence of sociolinguistic variables such as age, gender, and dialect.

## 5. Acknowledgements

Thanks to the NSF (award IIS-1216875). Thanks also to Jeremy Calder, Patrick Callier, Annette D'Onofrio, Katherine Hilton, Teresa Pratt, and Janneke Van Hofwegen for their helpful advice and thought-provoking discussion.

# 6. References

[1] Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. R package version, 1-0.

[2] Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [computer program]. Version 5.3. 39.

[3] Chafai, N. E., Pelachaud, C., Pel, D., & Breton, G. (2006). Gesture expressivity modulations in an ECA application. Intelligent Virtual Agents, 181-192.

[4] Chun, E. W. (2013). Ironic Blackness as Masculine Cool: Asian American Language and Authenticity on YouTube. Applied Linguistics, 34(5), 592-612.

[5] Cvejic, E., Kim, J., Davis, C., & Gibert, G. (2010). Prosody for the eyes: quantifying visual prosody using guided principal component analysis. Proceedings of INTERSPEECH 2010, 1433-1436.

[6] Gibbon, D. (2011). Modelling gesture as speech: a linguistic approach. Poznan Studies in Contemporary Linguistics, 47, 470.

[7] Gravano, A., Levitan, R., Willson, L., Benus, S., Hirschberg, J., & Nenkova, A. (2011). Acoustic and Prosodic Correlates of Social Behavior. Proceedings of INTERSPEECH 2011, 97-100.

[8] Guaïtella, I., Santi, S., Lagrue, B., & Cav, C. (2009). Are eyebrow movements linked to voice variations and turn-taking in dialogue? An experimental investigation. Language and Speech, 52(2-3), 207-222.

[9] Jannedy, S., & Mendoza-Denton, N. (2005). Structuring information through gesture and intonation. Interdisciplinary Studies on Information Structure, 3, 199-244.

[10] Kipp, M., Neff, M., & Albrecht, I. (2007). An Annotation Scheme for Conversational Gestures: How to economically capture timing and form. Language Resources and Evaluation, 41(3-4), 325-339.

[11] Krahmer, E., & Swerts, M. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. Journal of Memory and Language 57.3 (2007): 396-414.

[12] Kuznetsova, A., Brockhoff, P. B., & Christensen, R. (2013). lmerTest: tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package).

[13] Liscombe, J., Venditti, J., & Hirschberg, J. (2003). Classifying subject ratings of emotional speech using acoustic features. Proceedings of Eurospeech 2003.

[14] Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. Laboratory Phonology, 3(1), 71-89.

[15] Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. Journal of Artifial Intelligence Research, 30, 457-500.

[16] McFarland, D. A., Jurafsky, D., & Rawlings, C. (2013). Making the Connection: Social Bonding in Courtship Situations. American Journal of Sociology, 118(6), 1596-1649.

[17] McNeill, D. (1992). Hand and mind: What gestures reveal about thought. University of Chicago Press.

[18] Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility head movement improves auditory speech perception. Psychological Science, 15(2), 133-137.

[19] Pierrehumbert, J. B. (1980). The phonology and phonetics of English intonation. Doctoral dissertation, Massachusetts Institute of Technology.

[20] Rilliard, A., Shochi, T., Martin, J. C., Erickson, D., & Auberg, V. (2009). Multimodal indices to Japanese and French prosodically expressed social affects. Language and Speech, 52(2-3), 223-243.

[21] Scarborough, R., Keating, P., Mattys, S. L., Cho, T., & Alwan, A. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. Language and Speech, 52(2-3), 135-175.

[22] Walker, G. (2012). Coordination and interpretation of vocal and visible resources:Trail-offconjunctions. Language and Speech, 55(1), 141-163.

[23] Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., & Morency, L. (2013). YouTube Movie Reviews: In, Cross, and Open-domain Sentiment Analysis in an Audiovisual Context. Intelligent Systems, IEEE, 28(3), 46-53.