Voices of Vlogging

Joan-Isaac Biel and **Daniel Gatica-Perez**
Idiap Research Institute
École Polytechnique Fédérale de Lausanne (EPFL)
Switzerland
{jibiel,gatica}@idiap.ch

Abstract

Vlogs have rapidly evolved from the 'chat from your bedroom' format to a highly creative form of expression and communication. However, despite the high popularity of vlogging, automatic analysis of conversational vlogs have not been attempted in the literature. In this paper, we present a novel analysis of conversational vlogs based on the characterization of vloggers' nonverbal behavior. We investigate the use of four nonverbal cues extracted automatically from the audio channel to measure the behavior of vloggers and explore the relation to their degree of popularity and that of their videos. Our study is validated on over 2200 videos and 150 hours of data, and shows that one nonverbal cue (speaking time) is correlated with levels of popularity with a medium size effect.

Introduction

Vlogs are video collections that serve both as an audiovisual life documentary, and as a vehicle for communication and interaction on the Internet. Although vlogging is not exclusive of YouTube, the forms of social engagement inherent in vlogging are one of the key features that differentiate YouTube from a simple online video repository and distribution system to a platform for creativity and participation around video. Conversational vlogging is a predominant form of user generated content in YouTube, concentrating around 40% of the most viewed, discussed, favorited, and responded user created videos (Burgess and Green 2009).

Ethnographic studies have already investigated some of the underlying motivations, and the processes of creation and interaction through vlogging (Lange 2007; Molyneaux et al. 2008). However, despite the high popularity of vlogging, we do not know from any attempt to analyze conversational vlogs automatically. In addition, the analysis of the nonverbal behavior of vloggers has not been yet investigated. Nonverbal communicative cues (Knapp 2005) have shown to be robust and efficient descriptors to automatically measure human behavior in multiple human communication scenarios, such as individual face-to-face conversations or group meetings, and are consistent indicators of a number of attitudes and intentions of interacting people (Pentland 2008; Gatica-Perez 2009). In particular, the automatic analysis from "thin-slices" of nonverbal behavior has proven to be a good predictor of several social constructs and outcomes (Jayagopi et al. 2009), which is consistent with findings in social psychology, that establish that humans judgements based on first impressions are often correlated with subsequent assertions about people (Ambady and Rosenthal 1992). Recent research in social media has focused on the automatic analysis of the text content in personal websites and blogs to provide an insight on the personality traits and demographics of users based on their linguistic style (Gill, Nowson, and Oberlander 2009; Goswami, Sarkar, and Rustagi 2009). In addition, the concept of "thin-slices" has been mostly used to study the process of personality trait inference based on social networks' online profiles (Evans, Gosling, and Carroll 2008).

Our paper has two contributions. First, we present what to our knowledge is the first analysis of vlogs based on automatically extracted nonverbal behavioral cues. The nonverbal approach is relevant on its own and it opens a new domain for research, which is complementary to the existing text-based methods in the current literature. In this paper, we focus on four nonverbal cues extracted from the audio channel. Second, in a case study of the use of nonverbal behavior analysis in vlogs, we take popularity as an outcome of the interaction process in vlogging and we investigate how it relates to the nonverbal behavior of vloggers. Our study, validated on more than 2200 videos and 150 hours of audio, shows evidence that the nonverbal behavior of vloggers can be measured, and that some cues have a statistical relation to metadata that characterize their community of followers (viewers, commenters, or raters). Our work represents a first attempt towards the automatic analysis of vlogs, and in particular, towards the study of the nonverbal behavior of vloggers, which may lead to a better understanding of the processes involved in this new social communication scenario.

Data collection

We performed a manual annotation of YouTube videos to obtain a dataset of vlogs (YouTube's current taxonomy of videos and user channels doesn't offer any proxy for this purpose). We initially retrieved a set of 6335 videos from 878 users (containing up to 8 most recent videos per user) using the API and queries such as "vlog", "vlogging", "vlogger" on November 17th 2009. The annotation task, by means of a web-based tool, consisted on browsing all the

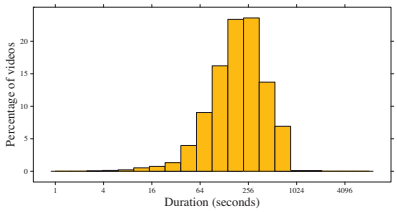

Figure 1: Histogram of the videos' duration in our dataset (median=202, mean=239.9, sd=165.6, max=2113 seconds).

videos from a given user, and answering few questions about them. We explicitly asked annotators to browse the videos using the progress bar, instead of watching them completely (typically, one person spend one hour to annotate the videos corresponding to 25 users). Based on these annotations, we identified a final set of 2269 videos from 469 users, which are categorized as *mainly conversational* vlogs with a single person (vlogger) talking on them and facing the camera addressing the audience during most of the time in a Skype-style fashion. We are interested in this setting as it represents the simplest vlogging scenario and the one that features most conversational behavior (compared to other vlogging styles that might feature music playing, mashups, etc.).

We collected all the videos together with their metadata, and extracted the audio track from them. Figure 1 shows the audio (video) length distribution in our dataset. Typical durations of vlogs are between 1 and 6min (70% of the videos appear in this interval), with a median duration of 3.4min. Only 2.4% of the videos are longer than 10min, a limitation that can be only exceeded by YouTube partners. This is equivalent to more than 7min of video per vlogger for 80% of the vloggers in the collection (only 6% have less than 2min), which is a reasonable amount of "thin-slice" data. Overall, our dataset contains 151 hours of video.

Metadata analysis

Statistics such as the number of views, comments, times marked as favorited, number of ratings, and average rating (all of them available from the videos' metadata) are measures of attention that reflect the patterns of engagement and participation around videos. Typically, the number of views has been taken as the main measure of video popularity, resembling the way audiences are measured in traditional mainstream media (Burgess and Green 2009). This metric tends to show the highest figures among all the metrics, not only because watching is the simplest passive form of participation in YouTube, but also because it does not require people to be registered or logged-in, which basically enables anyone with Internet access to watch a video. In contrast, the rest of the measures provide some account of popularity based on activities that signal a degree of participation from the YouTube audience, and are only allowed to registered users, which consequently results in lower values. In our dataset, for example, 10% of the videos had no ratings and 43% had not been marked as favorite any single time.

Figure 2 shows the metadata cumulative distribution for all the videos in our dataset. Except for the average rating, all distributions seem to exhibit power-law behavior (which is typically identified by a straight line in a log-log plot). This result is consistent with the analysis of other samples of YouTube videos, and it is a consequence of the preferential attachment mechanism that underlies the creation of incoming edges in many networks (Cha et al. 2007; Cheng, Dale, and Liu 2008). Specifically, what this indicates is that these popularity measures are not only measuring the videos' audience, but they also influence it, since they are used as a way to tell people what they should view, comment, favorite, or rate in the site. The cut-off on the distribution of "times favorited" may be an effect of the combination of a "fetch-at-most-once" behavior of users (Cha et al. 2007) with the lower popularity of this feature in YouTube. This is not reflected in the other distributions, because the most popular vlogs in our dataset have a smaller number of views (by one order of magnitude) than the most popular videos in the other samples referred in the papers above.

	1	2	3	4	5
1		0.80	0.93	0.75	0.88
2	0.80		0.86	0.59	0.79
3	0.93	0.86		0.64	0.93
4	0.75	0.59	0.64		0.64
5	0.88	0.79	0.93	0.64	

Table 1: Correlation between the videos' views (1), comments (2), ratings (3), favorite (4) and vloggers' subscribers (5) ($p < 10^{-5}$). The average rating showed no correlation.

In our sample, the views distribution is skewed towards a small number of views (median=231, mean=20030) with with 25% of the videos below 80 views (as an effect of the age of the videos). The number of comments (median=11, mean=284), ratings (median=9 mean=333.9), and faves (median=1, mean=60.4) are lower, with 25% of the vloggers having less than 3 comments, 3 ratings and no faves. Finally, average ratings are strongly biased towards high values (median=4.9, mean=4.6, and 60% of the videos having the average rate of 5), which may indicate that users tend to rate, often with very high scores, only when they liked the video, as a signal of sympathy with the content.

As shown in Table 1, these measures of user engagement and participation around videos show strong correlations between each other except for the average rating (likely due to its bias). More importantly, the videos' degree of popularity is strongly correlated to the vloggers' popularity, as measured by their number of subscribers.

Nonverbal cues extraction

We investigated four behavioral cues extracted from the audio channel that have been shown to be effective to characterize some constructs related to conversational interaction (e.g. interest, dominance, extraversion, or roles) both in the psychology literature (Knapp 2005) and more recently is social computing (Pentland 2008; Gatica-Perez 2009). While vlogs are obviously not face-to-face conversations, it is clear that vloggers often behave as if they were having a conversation with their audience (e.g. on Skype). These features are speaking energy, speaking time, speaking turns, and voicing rate, which, in rough terms, measure how loud, how much, how often, and how fast people speak, respectively (rather than understanding what they say). These cues have the advantage of being relatively easy to compute and robust. We automatically extracted the cues of each audio file using a modified version of the toolbox developed by the Human Dynamics group at MIT Media Lab (Pentland 2008).

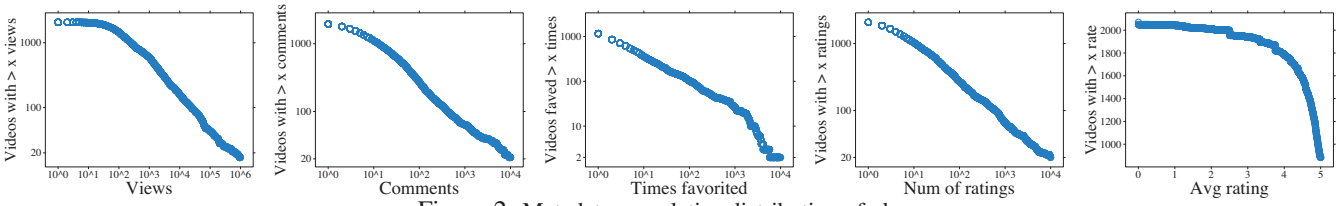

Figure 2: Metadata cumulative distribution of vlogs.

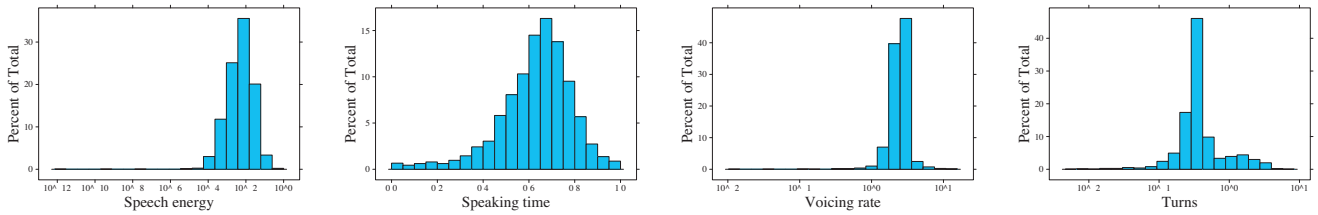

Figure 3: Nonverbal cues distribution of vlogs.

Speaking energy The speaking energy is a measure of emphasis, commonly considered a manifestation of dominant behavior. Confident, open, and influential people often speak louder, whereas lower values of energy may indicate a certain level or shyness and introversion (Knapp 2005). We segmented the audio track in speech/non-speech segments and computed the average energy of the speech signal.

Speaking time The speaking time is a measure of activity and dominance, usually correlated with interest level and extraversion in multi-party meetings (Jayagopi et al. 2009). In conversational vlogs, speaking time may vary from video to video and person to person, with some individuals being very talkative, and others adopting a more quiet behavior. We measured the fraction of time speaking dividing the total speaking time by the length of the video.

Voicing rate The voicing or articulation rate is a measure of how fast a speaker is producing phonemes during a burst of speech and it is a characteristic of the speaking rate of a person, i.e. how fast the person speaks (Basu 2002). Based on the division of speech segments between voiced and unvoiced regions, we divided the number of voicing regions by the total absolute speaking time.

Speaking turns A turn is defined as a single segment of speech existing between two pauses. In monologues, this cue is related to fluency (and pauses). We accumulated the total number of speaking turns over the entire video and divided it by the total absolute speaking time.

Figure 3 shows the distributions of the four features extracted for all the audio files. The speaking energy tends to be skewed towards low values (median=5.6e-3, mean=1.3e-2) and it exhibits a high variation for the whole sample (sd=0.025). The speaking time distribution (median=0.65, mean=0.64, sd=0.15) shows that 70% of the vloggers are speaking more than half of the time in the videos, which indicates that vloggers that were perceived as mainly talking in their vlogs are indeed making use of the floor a significant proportion of time. The voicing rate (median=mean=2.5, sd=0.6) varies between 2 and 4 regions per second, a range of values similar to other conversational scenarios (Morgan and Fosler-Lussier 1998). Finally, the speaking turns (median=0.34, mean=0.54, sd=0.6) exhibits very long tails, with 50% of the speakers concentrating betwen 0.28 and 0.42 turns per second, and the rest spreading from 0.008 to 4.

Nonverbal cues and popularity

Here we investigate whether these conversational cues can characterize vloggers with respect to their levels of popularity. We take popularity as a case study of an outcome of the interaction of vloggers with their audiences because it can be directly measured from their statistics. However, rather than accurately predicting popularity, our aim is to examine the validity of the proposed cues and the use of raw metadata statistics. First, we measure nonverbal behavior of vloggers in individual videos, and we relate it to the *videos*' degree of popularity (for reasons of space we present only correlations with the number of views, but other metadata showed similar performance). Second, by considering the videos as different "thin-slices" of the behavior of the same vlogger, we study the correlation between the aggregated behavior of the *vloggers* and their popularity, as measured by their number of subscribers.

Video popularity

We detected a key issue when investigating the relation between the nonverbal cues in videos and their number of views. The strong correlation between vloggers and videos popularity suggests that videos of different vloggers have intrinsically audiences of different sizes. This means that, independently of its content, a video uploaded to the channels of two vloggers with substantially different audiences is expected to receive a very different number of views (imagine the case of a very popular vlogger versus a non-popular one). More specifically, the attention received by a video is influenced both by the time spanned since published and the audience of a vlogger. Consequently, there is a need for an analytic strategy that discriminates these effects from those related to the video content itself (in our case, the nonverbal behavior). One possibility is to compute the correlations for subgroups of videos with similar ages and coming from vloggers with similar range of popularity. For our dataset, this would result in very small samples, so in practice, we only accounted for the temporal aspect. We sorted videos by their age and grouped them using overlapped varying-size temporal slices in order to obtain 20 samples of around 200 videos each. Figure 4 shows that for some of these samples there is a medium size effect correlation between

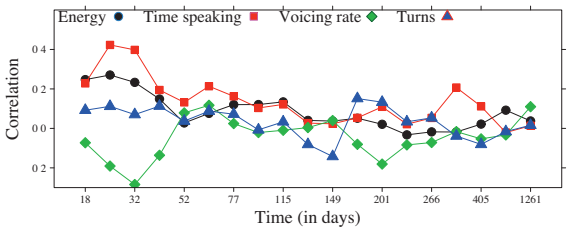

Figure 4: Correlation between nonverbal cues and videos' views for samples of videos of different age (the horizontal axis marks the age of the oldest video in each sample).

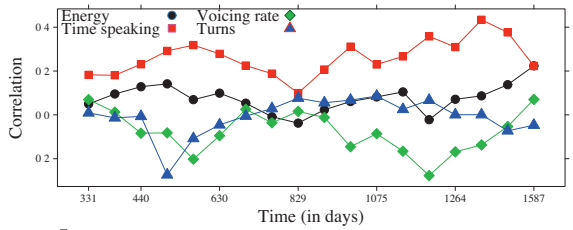

Figure 5: Correlation between aggregated nonverbal cues and vloggers' subscribers for samples of vloggers of different age (time since they joined YouTube, as taken from the user profile).

the videos' number of views and speaking energy ($r_{max} = 0.26, p < 10^{-3}$), speaking time ($r_{max} = 0.42, p < 10^{-8}$), and speaking turns ($r_{max} = -0.25, p < 10^{-5}$). No correlations were found for the voicing rate.

Vlogger popularity

We computed vloggers' nonverbal cues as the average cues over all their videos, and applied the above-mentioned strategy to obtain 20 samples of around 80 vloggers. Note that now, the temporal aspect (time since the vlogger joined YouTube) is the only effect for which to account. Figure 5 shows a medium size effect correlation between the popularity of vloggers and the speaking time ($r_{max} = 0.43, p < 10^{-3}$) that spans across all samples. Results for the other nonverbal cues are not significant ($p > 0.1$).

Discussion

We found that stronger correlations of nonverbal cues with measures of popularity were obtained in samples where there is a higher presence of popular vloggers, as compared to samples with lower correlations (results are consistent for both video and vlogger analysis). We hypothesize that the statistics of popular vloggers may be more consistent than those of non-popular ones as they result from aggregated behavior of larger audiences, which suggests that only in the first case the raw metadata statistics may be a reliable quantifier of the outcome of vlogging. This is coherent with the intuition that there may be users that are interested on interacting with small groups of people, rather than being followed by large audiences, which rises the question of where is the threshold between popular and non-popular vloggers.

Regarding the proposed nonverbal cues, we detected issues that may require refinements to measure nonverbal behavior in vlogs, and might lead to the consideration of alternative cues. For example, we noticed that the speaking energy could be capturing the volume of the microphone instead of the speaker emphasis, which does not happen in controlled scenarios. Instead, the SNR of the audio track could be used in the future.

Conclusions

We presented a first, original analysis of conversational vlogs from YouTube based on the automatic characterization of vloggers' nonverbal behavior. In a case study, we investigated two ways of using these cues to measure the nonverbal behavior of vloggers as it relates to their popularity and that of their videos. Our analysis show medium size effect correlations ($r \approx 0.30$) for three of the proposed nonverbal cues on different samples of our dataset. Although the observed correlations are somehow weak, our results disclose findings regarding the suitability of the proposed nonverbal audio cues, the reliability of using raw metadata statistics to measure the outcome of vlogging, and the challenge of finding proper analytical strategies to measure the effect of nonverbal behavior in vlogs. Future work may investigate on the use of nonverbal behavior to study other social psychology constructs related to vlogging, which will likely require human judgments complementary to the video metadata.

Acknowledgments We thank the support of the Swiss National Center of Competence (NCCR) on Interactive Multimodal Information Management (IM)2 and the voluntary annotators.

References

Ambady, N., and Rosenthal, R. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. *Psychological Bulletin* 111(2):256–274.

Basu, S. 2002. *Conversational scene analysis*. Ph.D. Dissertation.

Burgess, J., and Green, J. 2009. *YouTube: Online video and participatory culture*. Polity, Cambridge, UK.

Cha, M.; Kwak, H.; Rodriguez, P.; and Ahn, Y. Y. 2007. I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proc. of the 7th IMC*.

Cheng, X.; Dale, C.; and Liu, J. 2008. Statistics and social network of YouTube videos. In *Proc. of the 16th IWQoS*.

Evans, D. C.; Gosling, S. D.; and Carroll, A. 2008. What elements of an online social networking profile predict target-rater agreement in personality impressions. In *ICWSM 2008*.

Gatica-Perez, D. 2009. Automatic nonverbal analysis of social interaction in small groups: A review. *Image and Vision Computing* 27(12):1775–1787.

Gill, J. A.; Nowson, S.; and Oberlander, J. 2009. What are they blogging about? Personality, topic, and motivation in blogs. In *ICWSM 2009*.

Goswami, S.; Sarkar, S.; and Rustagi, M. 2009. Stylometric analysis of bloggers' age and gender. In *ICWSM 2009*.

Jayagopi, D. B.; Hung, H.; Yeo, C.; and Gatica-Perez, D. 2009. Modeling dominance in group conversations using nonverbal activity cues. *Trans. Audio, Speech and Lang. Proc.* 17(3):501–513.

Knapp, M. L. 2005. *Nonverbal communication in human interaction*. New York: Holt, Rinehart and Winston.

Lange, P. G. 2007. Publicly private and privately public: Social networking on YouTube. 1(13):361–380.

Molyneaux, H.; O'Donnell, S.; Gibson, K.; and Singer, J. 2008. Exploring the gender divide on YouTube: An analysis of the creation and reception of vlogs. *AC Journal* 10(2).

Morgan, N., and Fosler-Lussier, E. 1998. Combining multiple estimators of speaking rate. In *Proc. of ICASSP*.

Pentland, A. 2008. *Honest signals: How they shape our world*, volume 1 of *MIT Press Books*. The MIT Press.