2013-09-04

Judging Emotion from Low-pass Filtered Naturalistic Emotional Speech

John Snel
john.snel@student.dit.ie

Charlie Cullen
Technological University Dublin, charlie.cullen@tudublin.ie

Recommended Citation

Snel, J., Cullen, C. (2013) Judging Emotion from Low-pass Filtered Naturalistic Emotional Speech. Affective Computing and Intelligent Interaction (ACII), *Fifth biannual Humaine Association Conference on Affective Computing and Intelligent Interaction*, Geneva, Switzerland 2-5, September.

Funder: SFI

Judging Emotion from Low-pass Filtered Naturalistic Emotional Speech

John Snel
Digital Media Centre, Dublin Institute of Technology
Aungier St, Dublin 2, Ireland

Email: john.snel@mydit.ie

Charlie Cullen
Digital Media Centre, Dublin Institute of Technology
Aungier St, Dublin 2, Ireland

Email: charlie.cullen@dmc.dit.ie

***Abstract*—In speech, low frequency regions play a significant role in paralinguistic communication such as the conveyance of emotion or mood. The extent to which lower frequencies signify or contribute to affective speech is still an area for investigation. To investigate paralinguistic cues, and remove interference from linguistic cues, researchers can low-pass filter the speech signal on the assumption that certain acoustic cues characterizing affect are still discernible. Low-pass filtering is a practical technique to investigate paralinguistic phenomena, and is used here to investigate the inference of naturalistic emotional speech. This paper investigates how listeners perceive the level of Activation, and evaluate negative and positive levels, on low-pass filtered naturalistic speech, which has been developed through the use of Mood Inducing Procedures.**

***Keywords*-emotion in speech; low-pass filtering; Mood Inducing Procedures; emotional dimensions**

I. Introduction

Speech is an acoustically rich signal and comprises several constituent functions for communication: linguistic, paralinguistic, and extralinguistic. These three functions are an integral part of speech that are characterised by certain acoustical patterns. It has long been recognised that acoustical patterns, such as intonation, rhythm, and vocal intensity, signify paralinguistic cues that have communicative functions to express a person's affective state. The complex nature of spontaneous speech prohibits the complete separation of paralinguistic and linguistic cues in order to investigate one aspect independently. For speech that is truly 'natural' and 'spontaneous' in nature, it is impracticable to script its spoken dialogue to control for the linguistic content. Instead, researchers render speech incomprehensible by using masking techniques such as random splicing (e.g. [1]), backward speech (e.g. [2]), pitch inversion (e.g. [3]), and foreign speech (e.g. [4]). Inevitably, using these methods to mask linguistic content will affect certain acoustic properties that may characterise the emotion present e.g. backward speech will have reversed intonation contours.

Another masking technique used in paralanguistic studies [5], [6] is low-pass filtering. Low-pass filtering removes high frequencies which are important for speech comprehension (i.e. intelligibility), while leaving the lower frequency regions of speech intact. Tonal aspects of speech, which are dominated by the lower frequencies, play an important—if not most important—role in affect expression [7]. Low-pass filtering preserves prosodic characteristics, such as pitch parameters, stress patterns, rhythm, and tempo. An emotional judgement study that uses low-pass filtered speech, investigates how listeners judge and perceive emotion from these remaining prosodic features uninfluenced by the linguistic content.

The precise nature of low-pass filtering and its impact on speech perception remains to be established in speech intelligibility, and affect in speech studies. To our knowledge, few recent studies have used low-pass filtering as a masking technique on 'spontaneous' speech. One such study by McNally et al. [6] elicited emotion in participants—patients with panic disorder, major depressive disorder, social phobia, and healthy control participants—by asking them to recall both fear and neutral autobiographical memories. The speech clips were recorded onto audiotape, rather than being digitised, and content-filtered to eliminate frequencies above 400Hz. Each clip was evaluated by raters along the widely used scales: *negative*, *aroused*, and *dominant*. Two added scales, *anxious* and *sad*, were chosen applicable to the type of speech material being rated i.e. speech recordings from patients with mood and anxiety disorders. For the dimensions they studied, content-filtered speech conveyed enough information on fear related emotional valence. Similarly, Knoll et al. [5] studied perceptual ratings of vocal effect on filtered speech directed at Infants (IDS), Adults (ADS) and Foreigners (FDS). Raters were questioned on four scales: *positive* vocal affect, *negative* vocal effect, *encouragement of attention*, and *comforting and soothing*. The authors noted that certain affective scales may be more informative for a particular type of speech. Thus, comforting and soothing for example, might be more relevant for Infant Directed Speech (IDS). Four different filter conditions were investigated. It was acknowledged that cut-off frequencies above 1000Hz kept some semantic information discernible that may have confounding affects on raters' perception. The range that is important for speech intelligibility is about 500 to 5000Hz

[8]. However, there doesn't seem to be a standard optimal cut-off point for the level of intelligibility of filtered speech as it may vary depending on the proportion of background noise, quality of recording, and the speech characteristics of the speaker. To mask semantic content, researchers often use a *single* low-pass filter ranging from 300Hz-600Hz (e.g. [6], [9]). Alternatively, one can use a *set* of low-pass cut-off frequencies (e.g. [5]), or use a number of frequency bands called 'analysis filters' (e.g. [10], [11]). MacCallum et al. [12] recommend the cut-off frequency to be at least one octave above the F0 (minimum of 300Hz) to ensure acoustic analysis accuracy of percent jitter, percent shimmer, fundamental frequency (F0), signal-to-noise ratio (SNR), and nonlinear dynamic measures (correlation dimension and second-order entropy).

The aim of this experiment is to conceal any semantic information, and to compare ratings from its original condition with the ratings from its filtered condition. We do not apply a fixed cut-off value for all speech clips but rather we suggest a unique cut-off filter corresponding to every single clip. That is to say, for each clip we specify the cut-off point proportional to the clips key (F0 median). The speech dataset investigated here is based on naturalistic emotional content, for which elicited emotional content was achieved using Mood Inducing Procedures [13], [14]. Participants were asked to rate each speech clip in two conditions: (i) non-manipulated speech and (ii) filtered speech. Two emotion-scales were used for rating: Activation and Evaluation.

II. Methods

A. Design

As part of an emotional judgement task, participants were asked to take part in two separate listening tasks that were performed at least two weeks apart. For one task, they were asked to rate speech in its recorded form i.e. intact (non-filtered) and comprehensible, while for the other task they were asked to rate speech clips that were manipulated (low-pass filtered) to make them incomprehensible. A within-subject (repeated-measures) design was implemented, and in order to counterbalance carryover and order effects a crossover study was applied. In order to achieve this, subjects were (randomly) assigned to one of two groups so that the presented conditions were in different orders. That is to say, the first group rated the non-filtered speech on the first task and the filtered speech on the second task, and vice versa for the second group. The tasks were administered two weeks apart to reduce the subjects' retention of the speech tone (or F0) from the stimuli in the first task (testing effect). The speech stimuli were presented to the listener via a web-based rating tool. For each task, each participant was asked to rate 32 speech clips.

B. Experimental conditions

To deliver the task the rating tool was developed and hosted online. For this study, 57 participants completed the experiment. The participants were instructed to do the task using headphones in a quiet location to keep extraneous noise to a minimum. To emphasise the importance of performing in a quiet location, participants were asked to switch of TV's or radio's and minimise any other disturbances while doing the task.

C. Raters

All participants were fluent English, and anyone with known hearing impairments were excluded in this study. A total of 77 participants took part initially; however, only 57 completed the two phases of the experiment i.e. rating both conditions. Participants consisted of 30 males and 27 females with age ranging from 18 to 65 years.

D. Stimuli selection

The speech dataset used for this experiment was formed using task based Mood Induction Procedures (MIPs) [13], [14]. To capture emotional content, the researchers controlled and manipulated gameplay between participants in order to elicit emotional episodes. The induction experiments carried out were designed in such a way to capture the 'naturalness' of the elicited emotions, which incidentally produced subtle, underlying emotions.

An earlier study [15] collected ratings for this MIP speech dataset—using the same rating framework as in this paper—and concluded that the MIP procedures were successful in inducing non-neutral emotional states. From this existing dataset, we form the basis of the stimuli presented here for both conditions: non-filtered and filtered. A total of 64 speech clips (32 x 2 for each condition) were used.

1) Non-filtered stimuli: As part of a preliminary study for this experiment, we determined high inter-rater agreement values for each clip as measured by its standard deviation. Stimuli for this experiment were selected based on the high agreement values obtained. A total of 32 speech clips were selected; they were of short length ($\sim$5 seconds) assuming that no changes in emotion would occur.

2) Filtered stimuli: Because different speakers have different frequency ranges and spectrum energy distribution—including those for individual speech segments—using a fixed filtering condition across all speech clips could potentially give different levels of intelligibility. Aiming to make the level of unintelligibility uniform, it is suggested here to create a unique filter cut-off point that is proportional to the parameters of each speech clip. Each clip was filtered with a unique cut-off value proportional to its key (F0 median). The cut-off frequencies chosen were an octave above the

clip's key[1] (F0 median x 2), which ranged from 197Hz to 1162Hz for the 32 clips. All 32 clips were low pass-filtered (Hann window, smoothing=20, intensity=60.0) using Praat 5.3.13 [17] software.

3) Background information on filtering cut-off values: Two preliminary surveys, using two independent groups of 10 subjects, were carried out to (i) determine a suitable filtering condition to administer and (ii) to ensure that there was no comprehension of the spoken dialogue in the selected filtering condition.

For the first survey, we used 6 different speech clips for each of the 3 different low-pass filtering conditions. The three conditions were an octave above F0 min, key (F0 median), and F0 max. Unexpectedly, two clips were somewhat comprehensible in the filtered key condition but none in the F0 max condition; but as expected, participants could not comprehend any of the F0 min condition and remarked that the stimuli did not sound like speech, but just a 'rumbling noise'. We assessed that filtering an octave above the F0 min was excessively low and created inapplicable stimuli. Filtering an octave above the key is less likely to be incomprehensible compared to an octave above F0 max, so we chose the key value of the clip as the filtering reference point.

In the second survey, all 36 speech clips—chosen according to high inter-rater agreement—were low-pass filtered proportional to the speech clip's key. The 36 filtered clips were presented to 10 participants; some dialogue was correctly perceived in 4. These 4 clips were excluded from the main experiment, giving a final 32 speech clips for this experiment.

E. Web-based rating tool

The speech was rated on two discretised 5-point (colour-coded) scales: Activation and Evaluation. Great emphasis was put on the participants understanding of each scale. The tool includes a page with detailed instructions page about how and what to annotate. For each scale, the participant was provided with a definition and an accompanied written example. As mentioned above (see II-D1), inter-rater agreement values were obtained in a preliminary study. Based on these inter-rater agreement values, we provided an example speech clip for the instructions page to allow the participant to be fully conversant with each rating scale. This example was presented with its corresponding value on the Activation and Evaluation scale, which was determined by its Ground Truth value [18]—established from the same preliminary study[2] Each participant was required to successfully complete a multiple-choice questionnaire about the concept of Activation and Evaluation; subsequently, raters listened to each clip and rated accordingly. To avoid stimuli order effects, speech clips were randomised. Participants were given the option to skip a speech clip if they felt they could not rate it by choosing "Do not rate" but each clip could be replayed as many times as the participant wanted.

[1]To obtain the key (F0 median) value, we used a Praat script based on Celine De Loozes, 'Get_Speakers_register.praat'. This script minimises possible pitch tracking errors [16]. It can be found at: http://www.celinedelooze.com/MyHomePage/Praat.html.

	Non-filtered (α)	Filtered (α)
Activation	.588	.555
Evaluation	.26	.294

Table I: Krippendorff's α [19] as a measure for inter-rater reliability for both conditions on both scales.

III. Results

In this section, we present the results for the emotional judgement task for which we obtained data for both scales (Activation and Evaluation) in both conditions (non-filtered and filtered). As already mentioned, 57 participants took part that were asked to rate a total of 64 speech clips (32 for each condition). A total of 1823 ratings were received for the non-filtered speech clips (for Activation and Evaluation); only 1 DNR (Do not rate) was received, which amounts to only 0.05% of the total ratings received. This shows there was little uncertainty in the participants when rating the non-filtered speech clips. For the filtered speech clips, we received a total of 1815 ratings for each scale and 9 DNR ratings, which amounts to 0.49%. It would be expected that the filtered condition is harder to rate; however, this is also a small proportion of the total received ratings. We, therefore, disregard any further analysis on the DNR ratings and incorporate these values as 'missing values'.

For each scale, Figure 1 shows the number of ratings received for each class. We can observe from the Activation scale that the speech contains a relatively large number of active, non-neutral assets. The Evaluation scales shows that it contains a large number of neutral ratings, gradually decreasing towards Positive or Negative classes.

A. Inter-rater agreement

All 64 clips (32 for each condition) were rated by 57 raters. The clips that were not rated by participants are treated here as 'missing values'. We used Krippendorff's Alpha α [19] as a measure for inter-rater reliability. Krippendorff's Alpha α is a suitable measure to accommodate for missing values (DNR ratings), and where multiple raters are

[2]It should be noted that participants were informed that the provided example was based on the results from previous findings, and that the shown values did not necessarily indicate the correct chosen categories. Participants were informed of the values inconclusive nature.

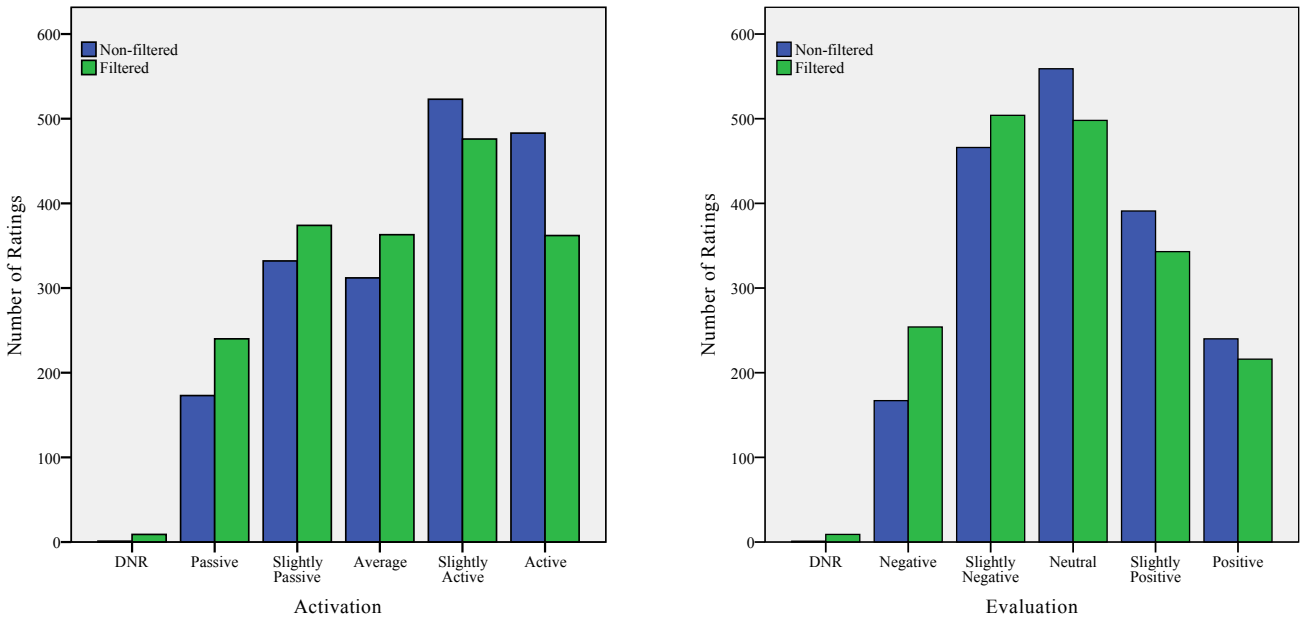

Figure 1: *Distribution of the ratings received for each condition, non-filtered and filtered speech—for the Activation (left) and Evaluation (right) scales. DNR = "Do Not Rate".*

	Non-filtered		Filtered	
Activation	Md	M	Md	M
Passive	1	4	1	5
Slightly Passive	6	7	9	7
Average	11	5	10	5
Slightly Active	5	6	6	9
Active	9	10	6	6

	Non-filtered		Filtered	
Evaluation	Md	M	Md	M
Negative	0	0	1	2
Slightly Negative	9	11	11	11
Neutral	15	13	13	9
Slightly Positive	7	3	4	6
Positive	1	5	3	4

Table II: The number of clips in each class with respective median (Md) and mode (M) values for the non-filtered and filtered conditions—for the Activation (left) and Evaluation (right) scales.

used. For each individual scale, Table I shows Krippendorff's Alpha α [19] (ordinal scale) for each stimuli condition. The agreement coefficients are higher for the Activation scale, with the highest α score observed on the Activation scale for the non-filtered condition. The observed alpha for Evaluation is slightly higher for the filtered condition. The overall results suggest low inter-rater reliability values.

B. Correlations of individual ratings between conditions

We calculated Kendall's τ_b between each rating for the non-filtered and filtered speech clips. The results for Activation (τ_b = .469, N = 1814, 2-tailed, $p < .0005$) indicate there was a moderate, positive correlation between the Activation received for the non-filtered speech and the filtered speech. In addition, we calculated the correlation for each participant on the Activation scale. Of the 57 participants, the correlation was small ($.1 < \tau_b \leq .3$) for 7 participants; for 19 participants the correlation was moderate ($.3 < \tau_b \leq .5$), and strong ($.5 < r \leq 1.00$) for the remaining 31 participants. For the analysis on Evaluation, the results for each rating (τ_b = .144, N = 1814, 2-tailed, $p < .0005$) show there was a small, positive correlation between the Evaluation perceived in the non-filtered speech and the filtered speech. The correlation obtained for each participant on the Evaluation scale showed that there was a small, negative correlation ($0 < \tau_b \leq$ -.3) for 14 participants, a small positive correlation for 31 participants (-.3 $< \tau_b \leq .0$), and a moderate, positive correlation ($.3 < \tau_b \leq .5$) for the remaining 12 participants.

C. Correlations of mean values for each clip between conditions

As well as measuring the correlation between individual ratings, we calculated Kendall's τ_b between the mean value for each clip of the non-filtered and filtered conditions (see

Fig 2). The mean values for each clip on the Activation scale (τ_b = .660, N = 32, 2-tailed, $p < .0005$) show that there was a strong, positive correlation between the mean values for Activation received for the non-filtered speech and the filtered speech. For the mean values on the Evaluation scale, the results for each rating (τ_b = .270, N = 32, 2-tailed, p = .31) show there was a small to moderate, positive correlation between the mean values for each clip on the Evaluation scale perceived in the non-filtered speech and the filtered speech. This shows that the correlation increased when the ratings were averaged for each speech clip.

D. Rating differences

The correlation coefficients showed association between the two conditions, but they are not affected by and do not demonstrate the overall differences in the ratings given in each condition. To investigate the differences in the perception of Activation and Evaluation between the two conditions (non-filtered and filtered), the ratings were subjected to the nonparametric Wilcoxon Signed Rank Test. The Wilcoxon Signed Rank Test revealed a statistically significant decrease in the level of Activation rated in the filtered condition compared to the non-filtered condition, z = -8.42, $p < .001$, with a small effect size (r = .14). The median Activation rating for the filtered clips (*Md* = 2 = Neutral) was lower than the Activation rating for the non-filtered clips (*Md* = 3 = Slightly Active). We can observe (Table II) that there are more clips with a median value for the *non-filtered* 'Active' class than there are for the *filtered* 'Active' class, but more instances in the *filtered* 'Passive' class than the *non-filtered* 'Passive' class. Similarly, we can observe that more instances of the mode value appear in the 'Passive' class, and 'Slightly Active' class.

For the Evaluation scale, we found a statistically significant decrease in the level of Evaluation perceived in the filtered condition, z = -4.833, $p < .001$. The effect size was small (r = 0.08). The median of Evaluation for the non-filtered clips (*Md* = 2 = Neutral) was the same for the overall median of Evaluation for the filtered clips (*Md* = 2 = Neutral). Table II shows that there are more instances of clips with median values in the extreme classes (1 Negative and 3 Positive) for the filtered condition. For the mode values in the filtered condition, the ratings occur more frequently in the 'Negative' and 'Slightly Positive' class, but less for the 'Neutral' and 'Positive' class.

IV. Discussion

We can observe from Figure 1 that non-filtered ratings skew more towards the 'Active' class for the Activation scale and the 'Positive' class for the Evaluation scale, which may be somewhat consistent with the assumption that upper frequencies lead to a loss of certain emotional cues [5]. However, for the filtered condition there is only a slight increase in the number of ratings received for the 'Average' class and, in fact, a slight decrease in the number of ratings received for the 'Neutral' class. For the Activation scale, there is an increase in the number of Passive ratings; similarly, there is an increase of Negative ratings for the Evaluation scale. This may suggest that the loss of upper frequencies, or linguistic content, is rated as more Negative and Passive.

Inter-rater reliability for the emotional judgement task was rather low for Activation in both conditions where α was .588 for the non-filtered condition and .555 for the filtered condition. More so, the Evaluation scale received α of .26 for the non-filtered condition and α of .294. This may be due to the difficulty of the task. First, rating spontaneous speech is a demanding task because the emotions are mostly underlying—milder and subtler than full-blown prototypical expressions. It is, after all, difficult to obtain natural speech with intense states through MIP experiments because of the restrictions on ethical matters. In saying that, much of the focus in recent studies is on underlying emotions that occur more frequently in day-to-day speech. Second, to some extent, the difficulty of the task may be related to the use of short speech clips ($\sim$5 seconds). However, the use of short clips was to minimise the possibility of emotional transitions and overlapping emotional states. In fact, research has suggested that participants can effectively recognise emotions as short as $\sim$5 seconds [13], [20].

The highest observed α score (.588) on the Activation scale for the non-filtered condition would be somewhat expected. The low α scores for Evaluation in both conditions show the difficulty of the task, whether or not the speech was filtered or not. It is clear that natural speech is inundated by ambiguity; in addition, the evaluator assesses emotional content in speech clip differently to how the emotion is truly felt by the speaker. Factors such as *display rules*, *deception*, and *systematic ambiguity* (see [21] for an overview) play an important role in the dissent of how emotion is perceived. Linguistic cues may not always concur with the paralinguistic cues; and, when evaluating speech, participants may prioritise acoustics over semantics, or vice versa. For this reason, it may be expected that the filtered condition would receive a higher inter-rater score than the non-filtered condition due to the minimisation of interference from the linguistic cues; this, however, was not the case.

The association of ratings between the two conditions was examined by calculating Kendall's τ_b over all individual cases— each individual rating of the non-filtered clip compared with the rating of the filtered clip. It showed that there was a moderate, positive correlation for the Activation scale, and a small, positive correlation for the Evaluation scale. The analysis on each participant showed the majority (31) of the participants had a strong correlation between each condition for the activation scale and a small positive correlation for

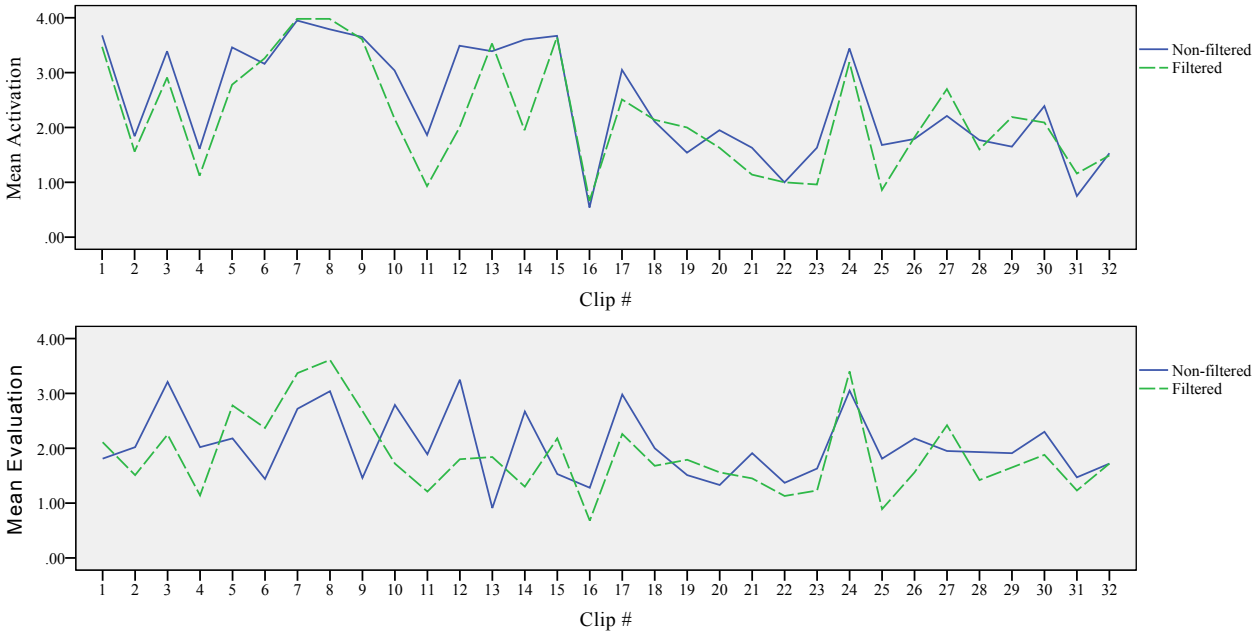

Figure 2: *Mean values obtained for each clip for Activation (above) and Evaluation (below). For the Evaluation scale, 0=Negative, 1= Slightly Negative, 2=Neutral, 3=Slightly Positive, 4= Positive. For the Activation scale: 0= Passive, 1=Slightly Passive, 2=Average, 3=Slightly Active, 4=Active.*

the evaluation scale. Interestingly, 14 participants showed a small, *negative* correlation between the two conditions for the evaluation scale. Again, this may be because the interpretation of paralinguistic cues may not correspond to the interpretation of linguistic cues. That is, emotions can be transmitted deceitfully, and speech that may be semantically negative may be expressed in a positive manner with, for example, laughter. In this case, linguistic and paralinguistic cues may have opposing impressions.

In addition to obtaining Kendall's τ_b between individual cases, we calculated Kendall's τ_b to compare the mean values for each clip in each condition (see Fig 2). The results indicated a strong, positive correlation for the activation scale and a small to moderate, positive correlation for the evaluation scale. This showed that the strength of correlation between the two conditions increased when comparing the mean values for each clip i.e. the correlations were stronger at the group level for each clip as opposed to the individual level of each rating—similar to the findings of Teshigawara et al. [1]. The strong correlation between the filtered and non-filtered conditions (see Fig 2) would be expected and in agreement with earlier findings (e.g. [22])—isolating pitch/prosody related features. The significant decrease in the perception level of activation, similarly expected, may be explained by the removal of the high frequencies. Clips 12 and 14 have the biggest difference (1.65 and 1.49 classes, respectively) between their mean activation values. Both speech clips contained laughter that may be a factor in the decrease in perception of activation level in the filtered condition. Although there is a small decrease in the level of evaluation for the filtered condition, the overall results do not necessarliy suggest a preference or importance in the emotion perception of lexical content, as there are fewer 'Neutral' ratings in the filtered condition. The low correlation, however, does suggest the incongruence between lexical and acoustical cues (cf. [23]). For the evaluation scale, clip 9 is rated more positive in the filtered condition compared to its counterpart in the non-filtered condition, its mean value is 1.46 in the non-filtered condition and 2.28 in the filtered condition; the spoken part in this clip: "oh my God... we were doing so well", may be semantically perceived as 'Slightly Negative', although it is rated 'Slightly Positive' in the filtered condition. This clip may be an example where the speech clip's *acoustic* significance opposes the affect of its *semantic* meaning.

V. Conclusions

In summary, the present study investigated the effect that low-pass filtering has on the perception of emotion—or more specifically Activation and Evaluation—in naturalistic speech. The results show that perception of Activation and Evaluation is influenced by low-pass filtering, and thus removing semantic content, but that it is relatively small. On that account, low-pass filtering is a useful tool to mask semantic content.

In future research, it would be of interest to obtain a speech dataset that provides for systematically ambiguous, and opposing acoustic and semantic meaning. This, however, may not be straight forward with naturalistic speech. With acted speech, however, one could systematically generate speech with opposing paralinguistic and linguistic content, such as speech with negative semantic content expressed with positive affect. It may be conceived that higher inter-rater agreement could be achieved for filtered speech of an ambiguous nature when its linguistic cues are removed.

VI. Acknowledgements

This work was supported by the Science Foundation Ireland under Grant No. 09-RFP-CMS253. Authors would like to express gratitudes to all raters, who participated in the research.

References

[1] M. Teshigawara, N. Amir, O. Amir, E. Wlosko, and M. Avivi, "Effects of random splicing on listeners' perceptions," *ICPhS*, 2007.

[2] M. J. Munro, T. M. Derwing, and C. S. Burgess, "Detection of nonnative speaker status from content-masked speech," *Speech Communication*, vol. 52, no. 7-8, pp. 626–637, Jul. 2010.

[3] K. R. Scherer, S. Feldstein, R. Bond, and R. Rosenthal, "Vocal Cues to Deception: A Comparative Channel Approach," *Journal of Psycholinguistic Research*, vol. 14, no. 4, pp. 409–425, 1985.

[4] J. Kim, "Bimodal emotion recognition using speech and physiological changes," *Robust Speech Recognition and Understanding*, pp. 265–280, 2007.

[5] M. A. Knoll, M. Uther, and A. Costall, "Effects of low-pass filtering on the judgment of vocal affect in speech directed to infants, adults and foreigners," *Speech Communication*, vol. 51, no. 3, pp. 210–216, Mar. 2009.

[6] R. J. McNally, M. W. Otto, and C. D. Hornig, "The voice of emotional memory: content-filtered speech in panic disorder, social phobia, and major depressive disorder." *Behaviour research and therapy*, vol. 39, no. 11, pp. 1329–37, Nov. 2001.

[7] K. R. Scherer, J. Koivumaki, and R. Rosenthal, "Minimal Cues in the Vocal Communication of Affect: Judging Emotions from Content-Masked Speech," *Journal of Psycholinguistic*, vol. 1, no. 3, pp. 269–285, 1972.

[8] D. A. Vickers, B. C. J. Moore, and T. Baer, "Effects of low-pass filtering on the intelligibility of speech in quiet for people with and without dead regions at high frequencies," *The Journal of the Acoustical Society of America*, vol. 110, no. 2, p. 1164, 2001.

[9] S. W. Gregory, W. Kalkhoff, S. K. Harkness, and J. L. Paull, "Targeted high and low speech frequency bands to right and left ears respectively improve task performance and perceived sociability in dyadic conversations." *Laterality*, vol. 14, no. 4, pp. 423–40, Jul. 2009.

[10] M. Ardoint and C. Lorenzi, "Effects of lowpass and highpass filtering on the intelligibility of speech based on temporal fine structure or envelope cues," *Hearing research*, vol. 260, no. 1-2, pp. 89–95, Feb. 2010.

[11] G. Gilbert and C. Lorenzi, "The ability of listeners to use recovered envelope cues from speech fine structure," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, p. 2438, 2006.

[12] J. K. MacCallum, A. E. Olszewski, Y. Zhang, and J. J. Jiang, "Effects of low-pass filtering on acoustic analysis of voice." *Journal of voice : official journal of the Voice Foundation*, vol. 25, no. 1, pp. 15–20, Jan. 2011.

[13] B. Vaughan, "Naturalistic Emotional Speech Corpora with Large Scale Emotional Dimension Ratings," Ph.D. dissertation, Dublin Institute of Technology, 2011.

[14] C. Cullen, B. Vaughan, S. Kousidis, Y. Wang, C. Mcdonnell, and D. Campbell, "Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction," *International Conference on Multidisciplinary Information Sciences and Technologies Extremadura (InSciT)*, 2006.

[15] J. Snel, A. Tarasov, C. Cullen, and S. J. Delany, "A Crowdsourcing Approach to Labelling a Mood Induced Speech Corpus," *4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals (ES^3 2012)*, 2012.

[16] C. De Looze and D. Hirst, "Detecting changes in key and range for the automatic modelling and coding of intonation," *Speech Prosody*, 2008.

[17] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (Version 5.3.47)," 2012. [Online]. Available: http://www.praat.org

[18] V. C. Raykar, S. Yu, L. H. Zhao, G. Hermosillo, C. Florin, L. Bogoni, and L. Moy, "Learning From Crowds," *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[19] K. Krippendorff, "Computing Krippendorff's Alpha-Reliability," *Departmental Papers (ASC)*, 2007.

[20] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[21] R. Cowie, E. Douglas-Cowie, N. Tsatatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion Recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[22] M. Schröder, "Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions," *Proc. Workshop on Affective Dialogue Systems Kloster Irsee, Germany*, vol. 209-220, 2004.

[23] L. C. Nygaard and J. S. Queen, "Communicating Emotion: Linking Affective Prosody and Word Meaning," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 34, no. 4, pp. 1017–1030, Aug. 2008.