Perception of Suspense in Live Football Commentaries from German and British Perspectives

Barbara Samlowski[1], Friederike Kern[2], Jürgen Trouvain[3]

[1]Amazon Development Center Germany (author was affiliated with [2] when this study was conducted)
[2]Faculty of Linguistics and Literary Studies, Bielefeld University, Germany
[3]Department of Language Science and Technology, Saarland University, Germany

samlowsk@amazon.de, friederike.kern@uni-bielefeld.de, trouvain@coli.uni-saarland.de

Abstract

The following study investigates the acoustic cues that British and German speakers rely on to assign the beginning of suspense during live football commentaries. In a perception experiment, participants were asked to listen to goal scenes broadcasted on British and German public radio, German private radio, and German public television, and to determine the point at which they felt that the suspense began. To disentangle textual cues from prosodic factors, a subgroup of participants were presented with delexicalized audio files that had been processed through a low-pass filter to eliminate any textual information, while a further group of participants based their decisions only on orthographic transcripts of the reporters' speech. The results indicated that British and German participants alike regarded a steep increase in fundamental frequency as a clear signal for the onset of suspense, while the verbal content of what was said played a subordinate role. However, there were also differences in the way in which suspense was perceived by German and English listeners. In particular, German participants were more consistent in their interpretation of delexicalized files than English participants, and did not gain as much from the additional information presented in the original files.

Index Terms: live sport commentaries, suspense, prosody, intonation, emotion perception

1. Introduction

Studies on the perception of emotion in speech have mainly focused on identifying specific emotion types rather than tracking the beginning of emotionally charged speech, e.g. [1, 2]. Even in automatic emotion recognition, where it may be crucial to recognize sudden changes in emotion, systems are trained and tested on individual utterances, each one marked with a specific emotion label, e.g. [3-5]. Perceptual changes in emotional nuances have been analyzed in research dealing with changes between non-emphatic and emphatic speech [6], the automatic tracing of emotions on a two-dimensional scale [7], the amount of speech it takes to recognize certain emotions [8], perceptual emotion changes in facial expressions [9], and the detection of emotion changes in speech [10]. Moreover, speech can reflect intentional affects and attitudes as well as spontaneous, uncontrolled emotions [14-16]. However, there are still many open questions with respect to how changes in emotions are expressed and perceived.

Live coverage of sports events in radio and television offer a rich field of highly emotional speech in a comparatively natural environment and as such tend to be less artificial than stimuli created for the sole purpose of analyzing emotion. Sports commentaries have been employed in several studies on the phonetic realization of excitement and suspense [11-13]. They show that it is the prosody that takes the important function to convey the suspense and excitement of the event to the audience.

To gain more information on how shifts from neutral to emotional speech are expressed and signaled, a perception experiment was conducted. Rather than identifying specific types of emotions, participants had to determine the starting point of emotional speech, in our case the beginning of suspense. One English and three German commentaries on goal scenes during a single football match were examined to investigate the extent to which listeners agreed on the onset of suspense during the narrations and what prosodic cues the reporters used to communicate the suspense to their audience. In the experiment, native German and English speakers were questioned in order to explore potential language differences in the perception of emotional suspense.

The paper presents a combination of descriptive statistical with qualitative analysis on differences in the ways original and manipulated sound files and transcript files were evaluated with regard to the beginning of suspense. The amount of acoustic material and the number of participants were too small to allow for inferential statistical methods.

2. Methods

2.1. Material

This study examined audio excerpts from four different live commentaries on the football match between Germany and England during the 2010 FIFA World Cup Round of 16 in South Africa. The commentaries were broadcasted on German public radio, British public radio, German private radio, and German public television. Unfortunately, we had no access to the commentaries of the British public radio and TV.

For reasons of comparison, the analysis focused on the six goal scenes featured in the match. Four goals were scored by the German team, while the English team scored one regular goal as well as one goal which was not given by the referee (goal '0' in Table 1). In each of the radio broadcasts, there were two commentators (one of whom on the German public radio station was female). The German television commentary was delivered by a single person. For each of the goal scenes in each of the commentaries, *original audio files* were extracted from about fifteen seconds before to five seconds after the start of the goal roar. Identification numbers for each of these twenty files are presented in Table 1.

Two additional sets of stimuli were created in order to examine how closely participants relied on prosodic cues as opposed to the verbal content of the spoken utterances. As a way of gaining judgments based exclusively on prosodic information, *delexicalized audio files* were created with low-pass band filters using the speech analysis software Praat [17]. Cut-

10.21437/SpeechProsody.2018-8

off frequencies of either 400 or 500 Hz were employed depending on the maximum pitch level of the audio file in question. Finally, to determine the impact of the verbal content alone, *orthographic transcriptions* were made of the reporters' utterances. The transcriptions did not contain any punctuation marks, information on stadium noise, or comments and interjections by the co-reporters.

Goal	Score	German Public Radio	British Public Radio	German Private Radio	German Public TV
1	1 – 0	111	121	131	211
2	2 – 0	112	122	132	212
3	2 – 1	113	123	133	213
0	not given	110	120	130	210
4	4 – 1	114	124	134	214
5	5 – 1	115	125	135	215

Table 1: *Identification numbers for the twenty-four audio files based on medium (radio vs. television), broadcasting station, and number of the goal*

Three hypotheses were formulated and followed up: First, it was assumed that pitch is the main factor for perceiving suspense; in this case, *original* and *delexicalized audio files* should lead to similar results. Second, we expected the perception of suspense to rely less on *what* was said rather than *how* it was said. This should lead to more variability in the evaluation of *orthographic transcriptions* than of original and delexicalized audio files. Finally, there should be no or little differences between English and German participants, considering the assumption that pitch is more important than verbal content for the determination of suspense beginning.

2.2. Participants

Thirty native speakers of German and twenty native speakers of British English took part in the study (different number due to time reasons). The German speakers were divided into three groups. The first ten participants were presented with the original audio files, the second ten with the delexicalized audio files, and the third ten with the orthographic transcriptions of only the German commentaries.

The English speakers were divided into two groups of ten each. Participants in one group were given the original audio files, while those in the other group were first presented with the delexicalized files and then with the orthographic transcriptions of only the English commentaries (the transcripts could not be linked to the previously linked delexicalized audio files). Thus, *six evaluation groups* were established. Most of the German participants rated their knowledge of English as medium to good, whereas the most of the English participants reported that they had little to no knowledge of German. This unavoidable situation regarding language proficiency led to a misbalance for the audio files.

2.3. Procedure

A Praat script [17] was used to collect estimations of the starting point of suspense in the original and delexicalized audio files. Participants listened to the audio files in randomized order over headphones. Each file was visually represented by a depiction of the sound waveform and a TextGrid tier. Participants were allowed to listen to the audio files multiple times, tracking the vertical line across the sound waveform to determine at which point during the file the suspense began. Depending on their familiarity with the program, they could either set the boundary themselves or indicate where they felt the boundary should be to the experimenter, who then set the boundary for them. Care was taken to ensure that experienced users of the program did not zoom into the audio file or use options such as pitch tracking to help them in their decision.

The orthographic transcriptions of the commentaries were given to the participants with the instruction to mark the starting point of suspense in each transcription with a vertical line. For the results the text-based starting points of suspense were aligned with the audio files so as to make them comparable to the starting points perceived in the original and delexicalized audio files. Participants were generally allowed to place a second boundary in a file if they noticed two separate incidents of suspense. If no suspense was perceived at all, the boundary was placed at the very end of the file.

3. Results

Starting points of perceived suspense in the twenty-four recordings with goal scenes were compared within and across the six groups, namely German and English interpretations of original audio files, delexicalized audio files, and orthographic transcriptions. The audio files were also examined acoustically to identify possible prosodic and textual cues of suspense which might be able to explain some of the variation in the perceived starting points of suspense. Data analysis and visualization were performed in R [18]. Descriptive statistical analysis was employed to provide an overview over tendencies of more or less variability within and across groups; detailed auditory and acoustic phonetic analysis of single files was then employed to explain the findings.

3.1. Focus on Variability within Evaluation Groups

To determine the level of agreement among the participants in each of the six evaluation groups, interquartile ranges of the indicated time points were calculated for each of the twenty-four audio files. In addition, an acoustical analysis was performed on the files with the five largest differences in interquartile ranges across evaluation groups to determine possible reasons for discrepancies between the interpretation of original and delexicalized audio files in the two languages. Figure 1 is a boxplot diagram of interquartile ranges for English and German participants confronted with the three sets of stimuli.

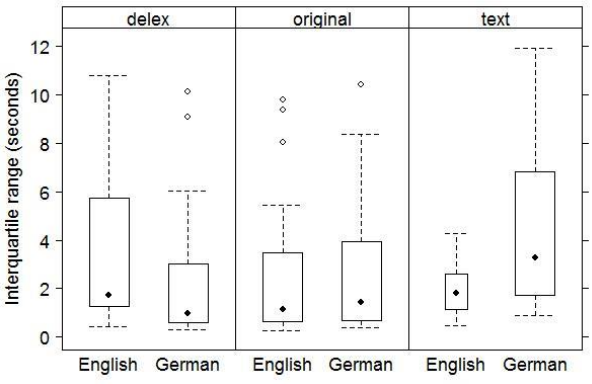

Figure 1: *Interquartile ranges of perceived suspense onsets by participants' language and stimulus type*

English evaluations of *delexicalized audio files* were considerably more variable than *German* evaluations of the same files. Since the files contain no verbal information, differences between these groups would only be expected if German and English speakers used different prosodic cues to encode suspense. The results indicate that the English

participants may have been less adept than German participants at decoding information on pitch in the predominantly German files. For instance, in the delexicalized version of File 130, English participants seemed to pay attention to strong phrasal pitch accents during the early part of the commentary to a greater extent than German speakers who mainly regarded the steep increase in pitch level before the goal as a sign of beginning suspense. In cases with a gradual increase in pitch level, English participants showed no clear evaluation pattern for delexicalized files, while German participants focused on one particular point in the pitch curve (Files 122, 124, 214).

English participants were more consistent when annotating *original files* than when interpreting *delexicalized data*. They may have been able to make use of additional information in the original files, such as stadium noise (Files 112, 213) or interjections by co-reporters (Files 124). However, such cues did not consistently decrease the degree of variability among annotators. In file 215, most English annotators seemed to consider either a change in voice quality at the beginning of the file or the final pitch upstep before the goal roar as indicators of suspense. In contrast, annotations for the *delexicalized files* tended to be assigned between these two points, resulting in higher interquartile ranges for *original* compared to *delexicalized* file evaluations.

German participants were comparatively consistent in determining the onset of suspense in the *delexicalized files*, and the additional information provided in the *original files* did not lead to a general increase in their consistency. Only in some cases did information concerning the verbal content help participants to narrow down the perceived onset of suspense (Files 112, 130). However, there was an overall tendency for German interpretations of *original files* to have slightly larger interquartile ranges than of *delexicalized files*, or *English* interpretations of *original files*. The results suggest that while German participants sometimes paid attention to other suspense indicators, such as voice quality (File 214), interjections by co-reporters (Files 122, 124) or verbal information (File 110), they were also strongly influenced by prosodic cues such as sudden pitch upsteps (Files 110, 214) or phrasal intonation contours with a high tonal register, very little pitch variation and no final falling (Files 122, 124).

As expected, *German* evaluations based on *orthographic transcriptions* tended to be less consistent than evaluations of original or delexicalized files. *English* speakers, on the other hand, were comparatively consistent when determining the onset of suspense on the basis of the verbal content alone. However, an investigation of individual files showed that for the first three goals (Files 121, 122, and 123), evaluations of the orthographic transcriptions by English annotators were actually less consistent than annotations in any of the other evaluation groups. Only for Files 124 and 125 did the interpretation of orthographic transcriptions by English annotators lead to lower interquartile ranges when interpreting the orthographic transcription than when annotating original or delexicalized audio files. In these cases, most of the English participants rated the phrases "it's Schweinsteiger" (File 124) and "and he's away" (File 125) as indicators of suspense beginning. Only minimal variability was found with regard to the goal not given by the referee (File 120).

3.2. Focus on Variability across Evaluation Groups

Median values for the onset of suspense as perceived by the six evaluation groups were calculated and compared for each of the twenty-four goal scene commentaries (see Figure 2).

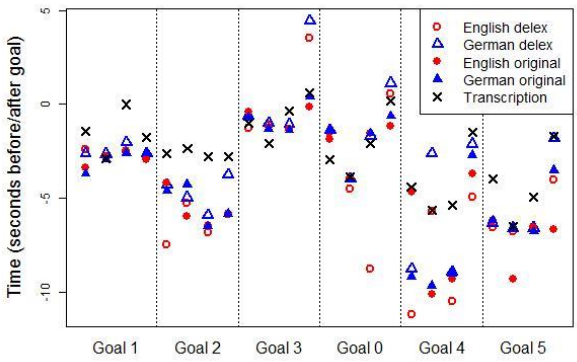

Figure 2: *Median values of perceived suspense onsets for each goal scene (left to right in each goal section: German public radio, British public radio, German private radio, German publ. TV). 'English delex' refers to English listeners of delexicalized versions, etc.*

Since English and German participants evaluated only orthographic transcriptions in their own language, English and German text based evaluations ('transcription') are represented by the same symbol in this diagram. English text evaluations can be found in the second column for each goal (British public radio), while the first, third, and fourth columns (German public and private radio, German public television) contain text evaluations by German annotators.

For most of the commentaries, median values for interpretations based on *audio files* were similar to each other, while noticeable differences appeared in comparisons with evaluations based on *orthographic transcriptions*. Even in Files 124 and 125, where English participants showed a high level of agreement in the interpretation of the orthographic transcriptions due to the specific verbal information given (see above 3.1), their perception of suspense based exclusively on the verbal content bore little resemblance to their assessment of the original audio files.

In a few cases, there were strong differences between interpretations of *delexicalized files* compared to the *other files*. *English* speakers tended to assign comparatively early starting points. This may be due to their interpretation of the animated 'speech style' [12] at the beginning of the files as a sign of suspense (Files 112, 114, 130). Evaluations of delexicalized files by *German* speakers tended to occur at comparatively later time points (Files 124, 212, 213, 215). In the *television commentaries* on the goals for the English team (Files 213, 210), German and English speakers tended to perceive suspense in the delexicalized files only after the goal had occurred. A number of German speakers did not perceive any suspense in the television commentaries, placing the boundary at the very end of the file.

All four commentaries of the fourth goal elicited large differences between individual evaluation groups. Here, there tended to be two potential starting points for perceived suspense. While all files showed a slight increase in pitch immediately before the goal itself, suspense was also perceived comparatively early during the goal scene due to pitch increases (Files 114, 134), textual cues and interjections by the co-reporter (File 124), or a rough voice quality (File 214). During commentaries on the fifth goal, many English participants assigned particularly early starting points of suspense in the original files in reaction to an interjection by the co-reporter (File 125) or a rough voice quality of the speaker (File 215).

3.3. Pitch and Variability

Investigations of variability within and across evaluation groups revealed that the beginning of suspense was generally more ambiguous for some files than for others. An acoustic analysis of the individual audio files showed that almost all files which elicited a high level of agreement between participants featured a steep increase in pitch at the point where most starting points were placed (see, for instance, Figure 3). Suspense boundaries were especially close together when the increase was followed by a horizontal or slightly declining pitch contour which clearly marked the end of the increase. Files with a more gradual pitch movement or two separate pitch increases tended to elicit more variation of perceived onsets of suspense within and across evaluation groups.

4. Discussion and Conclusions

This study aimed to investigate three hypotheses concerning the perception of suspense in original and delexicalized audio files by German and English participants. First, assuming that pitch is the main factor for the perception of suspense (similar to goal roar in even more extreme pitch registers [19]), then original and delexicalized data should lead to similar perceptual results. While this was predominantly the case for the *German* participants, *English* participants often disagreed on where suspense started in the delexicalized files and in some cases tended to perceive suspense at an earlier point in time than the other groups. Even though the English participants knew little to no German, they showed considerably less variability in assessing the *original files* in comparison with the *delexicalized files*. This suggests that while English participants may have been comparatively less adept at interpreting the (predominantly German) commentaries merely on the basis of prosodic cues, they were probably able to make use of information such as voice quality. It remains unclear whether stadium noise or interjections by co-reporters influenced the determination the onset of suspense. Only in two cases did their interpretations of the original data lead to markedly different results than those of German speakers.

Second, if suspense perception is based less on *what* is said and more on *how* something is said, suspense boundaries in orthographic transcriptions should occur at different time points and show a larger variability than in original or delexicalized audio files. The results suggest that the influence of verbal content on the perception of suspense is, in fact, minimal. Although for two of the commentaries English participants assigned comparatively clear onsets of suspense based on orthographic transcriptions alone, the boundary placement in the text files did not correspond to English interpretations of the original audio files.

Third, if suspense is encoded similarly in English and German, and if the verbal content of what is said is irrelevant, there should be no differences between English and German interpretations of the audio files. Both English and German participants did perceive a steep pitch rise as a clear sign of the beginning of suspense. However, while German speakers tended to attach more importance to pitch increases and were often able to uniformly latch on to a certain point during the pitch contour even in delexicalized files with no salient pitch rise, English speakers were less consistent in their estimations of delexicalized files and more inclined to regard strong phrasal pitch accents resulting in a wide pitch range as a sign of suspense. When examining original sound files, English participants seemed to pay more attention to voice quality and background events than German participants, who were often divided over whether or not to prioritize these cues over their interpretation of the pitch contour.

In conclusion, the study showed that *what* is said during football commentaries is not as important for the perception of suspense as *how* it is said. For German as well as English speakers, a sudden sharp increase in pitch was found to be a clear indication of suspense. English participants were found to be less consistent in their interpretation of delexicalized files than German participants. This may suggest that English speakers generally rely more on voice quality or background utterances and noise to interpret suspense. On the other hand, it may also be the case that the two languages use pitch contours in slightly different ways and that the English participants found the predominantly German contours difficult to interpret.

Future studies should include a more balanced data set, ideally more data, and also other speaking situations with suspense. On the perceptual side it would be worth to check whether any kind of strong change in the signal would be a good cue for getting the impression of start of suspense. Also a comparison of less dramatic with more dramatic scenes would be a good point to continue the research on suspense detection.

5. Acknowledgements

This study was funded by the German research foundation DFG. Many thanks to Sascha Schäfer for recruiting the native English speakers and conducting the perception experiments with them. The authors thank two anonymous reviewers for their comments.

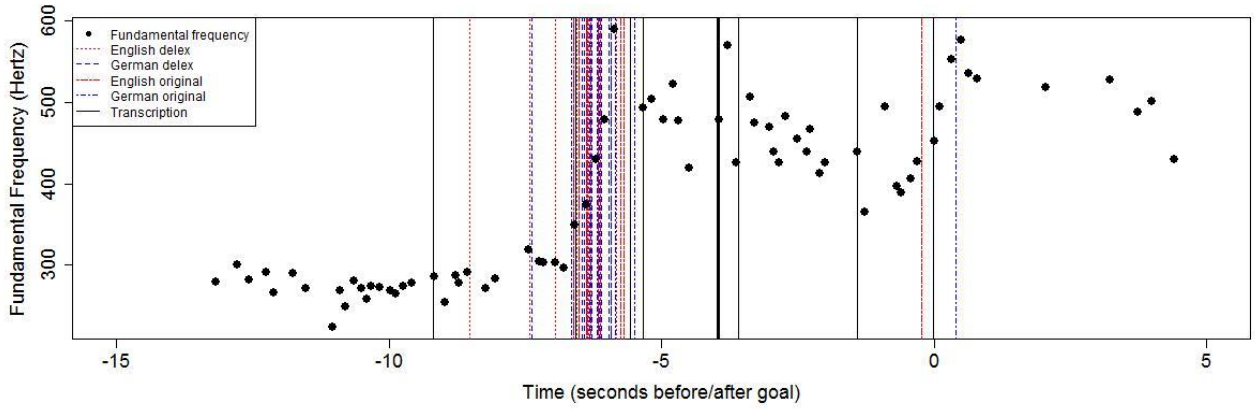

Figure 3: *Pitch contour for File 115 (female speaker, median F0 per syllable nucleus) with 43 perceived suspense boundaries*

6. References

[1] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614-636, 1996.

[2] T. Bänziger and K. R. Scherer, "The role of intonation in emotional expressions," *Speech Communication*, vol. 46, pp. 252-267, 2005.

[3] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately seeking emotions or: Actors, wizards, and human beings," in *ITRW Speech and Emotion - ISCA Tutorial and Research Workshop on Speech and Emotion, September 5-7, Newcastle, Northern Ireland, Proceedings*, 2000, pp. 195-200.

[4] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *ICSLP 2002 - 7th International Conference on Spoken Language Processing, September 16-20, Denver, Colorado, Proceedings*, 2002, pp. 2037-2040.

[5] M. Grimm, K. Kroschel, H. Harris, C. Nass, B. Schuller, G. Rigoll, and T. Moosmayr, "On the necessity and feasibility of detecting a driver's emotional state while driving," in *Affective Computing and Intelligent Interaction - Second International Conference ACII 2007 Lisbon, Portugal, September 2007 Proceedings*, A. C. R. Paiva, R. Prada, and R. W. Picard, Eds., Berlin: Springer Verlag, 2007, pp. 126-138.

[6] M. Selting, "Emphatic speech style - with special focus on the prosodic signalling of heightened emotive involvement in conversation," *Journal of Pragmatics*, vol. 22, no. 3/4, pp. 375-408, 1994.

[7] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, '"Feeltrace': An instrument for recording perceived emotion in real time," in *ITRW Speech and Emotion - ISCA Tutorial and Research Workshop on Speech and Emotion, September 5-7, Newcastle, Northern Ireland, Proceedings*, 2000, pp. 19-24.

[8] M. D. Pell and S. A. Kotz, "On the time course of vocal emotion recognition," *PLoS ONE*, vol. 6, no. 11, p. e27256, 2011.

[9] V. Sacharin, D. Sander, and K. R. Scherer, "The perception of changing emotion expressions," *Cognition and Emotion*, vol. 26, no. 7, pp. 1273-1300, 2012.

[10] C. N. v. d. Wal and W. Kowalczyk, "Detecting changing emotions in human speech by machine and humans," *Applied Intelligence*, vol. 39, no. 4, pp. 675-691, 2013.

[11] J. Trouvain and W. J. Barry, "The prosody of excitement in horse race commentaries," in *ITRW Speech and Emotion - ISCA Tutorial and Research Workshop on Speech and Emotion, September 5-7, Newcastle, Northern Ireland, Proceedings*, 2000, pp. 86-91.

[12] F. Kern, "Speaking dramatically: The prosody of live radio commentary of football matches," in *Prosody in Interaction*, D. Barth-Weingarten, E. Reber, and M. Selting, Eds., Amsterdam; Philadelphia (PA): John Benjamin's Publishing Co., 2010, pp. 217-238.

[13] S. Audrit, T. Pršir, A. Auchlin, and J.-P. Goldman, "Sport in the media: a contrasted study of three sport live media reports with semi-automatic tools," in *Speech Prosody 2012 - Sixth International Conference on Speech Prosody, May 22-25, Shanghai, China, Proceedings*, 2012, pp. 127-130.

[14] A. Rilliard, T. Shochi, J.-C. Martin, D. Erickson, and V. Aubergé, "Multimodal indices to Japanese and French prosodically expressed social affects," *Language and Speech*, vol. 52, no. 2/3, pp. 223-243, 2009.

[15] I. Grichkovtsova, M. Morel, and A. Lacheret, "The role of voice quality and prosodic contour in affective speech perception," *Speech Communication*, vol. 54, pp. 414-429, 2012.

[16] A. Hönemann, H. Mixdorff, and A. Rilliard, "Social attitudes - Recordings and evaluation of an audio-visual corpus in German," in *Forum Acusticum 2014 - Seventh Forum Acusticum, 9-12 September, Kraków, Proceedings*, 2014.

[17] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341-345, 2001.

[18] R Core Team, "R: A language and environment for statistical computing," ver. 3.1.1. Vienna, Austria: R Foundation for Statistical Computing, 2014. Available: http://www.r-project.org/.

[19] Trouvain, J, "Between excitement and triumph — live football commentaries in radio vs. TV" in Proc. *17th International Congress of Phonetic Sciences*, Hong Kong, pp. 2022-2025, 2011.