2008-6

A VOWEL-STRESS EMOTIONAL SPEECH ANALYSIS METHOD

Charlie Cullen
Dublin Institute of Technology, charlie.cullen@dit.ie

Brian Vaughan
Dublin Institute of Technology, brian.vaughan@dit.ie

Spyros Kousidis
Dublin Institute of Technology, spyros.kousidis@dit.ie

Recommended Citation

Cullen, C., Vaughan, B. & Kousidis, S. (2008) A VOWEL-STRESS EMOTIONAL SPEECH ANALYSIS METHOD, *5th International Conference on Cybernetics and Information Technologies, Systems and Applications,* University College Dublin, 29 Jun-2 July.

A VOWEL-STRESS EMOTIONAL SPEECH ANALYSIS METHOD

Charlie Cullen, Brian Vaughan* and Spyros Kousidis**

***Cognition Speech & Audio Lab, Digital Media Centre, Dublin Institute of Technology, Aungier Street, Dublin 2, Ireland**

ABSTRACT

The analysis of speech, particularly for emotional content, is an open area of current research. This paper documents the development of a vowel-stress analysis framework for emotional speech, which is intended to provide suitable assessment of the assets obtained in terms of their prosodic attributes. The consideration of different levels of vowel-stress provides means by which the salient points of a signal may be analysed in terms of their overall priority to the listener. The prosodic attributes of these events can thus be assessed in terms of their overall significance, in an effort to provide a means of categorising the acoustic correlates of emotional speech. The use of vowel-stress is performed in conjunction with the definition of pitch and intensity contours, alongside other micro-prosodic information relating to voice quality.

Keywords— Acoustic signal analysis, Speech analysis, Speech processing, Speech Corpus.

1. INTRODUCTION

Existing work in the field of emotional speech research has considered means by which suitable speech assets may be obtained [1-3], leading to the creation of a corpus of natural emotional speech [4-6]. Circumplex emotional modelling [7-9] is then used to rate emotional speech assets on unit scales of activation and evaluation (Figure 1):

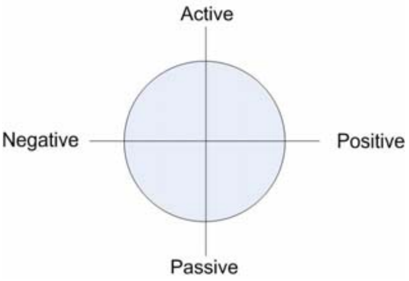


Figure 1: Circumplex emotional model denoting dimensions of activation and evaluation, adapted from Scherer [7]

Analysis of emotional speech assets is thus performed using defined emotional parameters, so that any potential acoustic correlates of emotional speech can be measured in relation to them. The corpus can then be used to determine the acoustic correlates of emotional speech by comparing the analysis results of speech assets. Listening tests in conjunction with dimensional rating will be used to create a statisitcal analysis of the speech corpus [5]. Thus the corpus can be defined in terms of its emotional content prior to acoustic analysis using the vowel-stress analysis framework proposed in this paper.

2. STRESS AND PROSODY IN SPEECH

Prosody in speech can be considered in many different ways [10, 11], from purely linguistic analysis which places little focus on the acoustic elements of speech to a supra-segmental approach including pitch, loudness and speech rate [12, 13]. To facilitate comparison, Dutoit [14] suggests 3 different representations of prosody based on acoustic, perceptual and linguistic attributes (Table 1):

Acoustic	**Perceptual**	**Linguistic**
Fundamental Frequency (F0)	Pitch	Tone, intonation, aspect of stress
Amplitude, Energy, Intensity	Loudness	Aspect of stress
Duration	Length	Aspect of stress
Amplitude dynamics	Strength	Aspect of stress

Table 1: A comparison of acoustic, perceptual and linguistic representations of prosody, adapted from Dutoit [14].

The acoustic representation refers to measurable acoustic properties of the speech signal such as fundamental frequency (F0), amplitude and duration (vowel duration, pitch duration, intensity duration, etc). It is important to note that acoustic representations of prosody are context sensitive, referring to speech events from syllable and word level through to longer sentence and clip events.

The perceptual representation of prosody is largely the same as the acoustic, though it defines differing terms for the same acoustic properties (pitch, loudness, length and strength). A more general description is taken by the linguistic approach, which defines all attributes as aspects of linguistic stress. From the linguistic perspective stress is used to distinguish between an emphasised phrase in a sentence or an individual stressed word. Having said this, there is no consensus as to the definition (or scope) of linguistic stress [15], particularly as it relates to acoustic parameters [14].

3. VOWEL STRESS ANALYSIS

The isolation of vowel events in a speech signal is a common approach in speech analysis [16-19]. Additional argument is provided by cognitive studies of infant language perception [20-22], that indicate a preference for vowels rather than consonants [23]. Although a syllable may be formed around non-vocalic events, most speech patterns involve the alternation of vowels and consonants [24]. As a result, many approaches to speech analysis have considered the use of vowel, consonant vowel (CV) and consonant-vowel-consonant (CVC) structures [19, 24, 25] for automatic rhythm extraction. The definition of the ‘pseudo-syllable’ [25] is based on the observation that the CV structure is the most common structure [19, 26], and thus leads to the use of a vowel onset detection algorithm to determine the occurrence of each vowel (and hence each CV) in a speech event. To perform acoustic analysis of emotional speech events, an application called LinguaTag [27] was constructed (Figure 2):

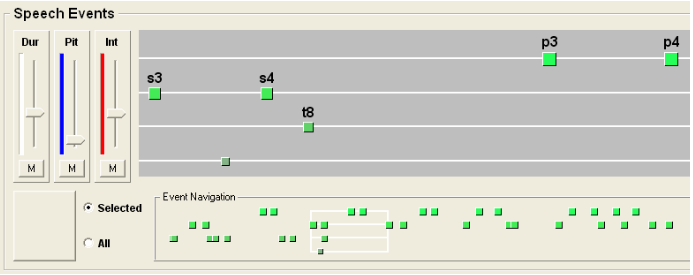


Figure 2: Screenshot of vowel stress events in the LinguaTag speech analysis application [27]

LinguaTag uses the Praat analysis engine [28] to obtain acoustic information about vowel events in a speech signal, which are then displayed in the application GUI for acoustic, linguistic and emotional analysis prior to output of this information in SMIL file format [29].

4. LINGUISTIC PRIORITISATION OF VOWEL STRESS EVENTS

One aspect of the proposed method is the grading or prioritisation of vowel stress events, to facilitate hierarchical analysis of those events. This research considers the definition of a linguistic foot [30] to determine the location of different levels of stress within a spoken utterance. Using this concept of the foot to define spoken rhythm, the structuralist method [16, 21] defines four levels of rhythmic stress (or foot) within linguistics; primary, secondary, tertiary and weak (Figure 3):

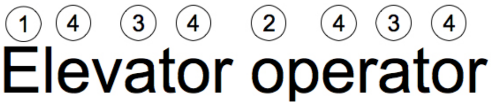


Figure 3: Example of structuralist rhythmic stress definitions

This method of graded prioritisation of linguistic stress (though limited in its current linguistic application [31]) may possibly indicate a means of grading the acoustic parameters related to stress. It is argued that although a direct correlation between acoustic parameters and linguistic stress may not be possible without recourse to lexical complexities, it may be possible to employ a method of prioritisation for the purposes of acoustic analysis of vowel stresses in a speech signal. With this in mind, this paper proposes a method of stress prioritisation based on three fundamental acoustic attributes of a vowel event: pitch, intensity and duration.

5. VOWEL STRESS DEFINITION

The acoustic attributes of pitch, intensity and vowel duration were chosen as fundamental features of a speech event, which are agreed as being common to all 3 representations models of speech [14]. The determination of a stress is therefore based on these 3 parameters, by determining the variation from the mean of each (Figure 4):

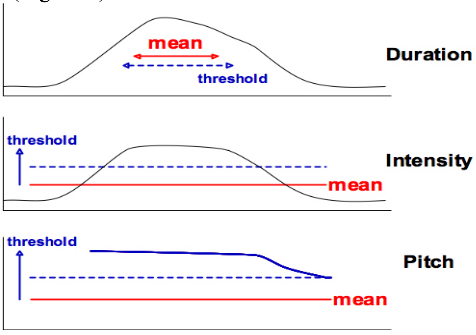


Figure 4: Definition of vowel threshold levels

A vowel event can thus be graded in terms of threshold values relative to those means (Figure 5):

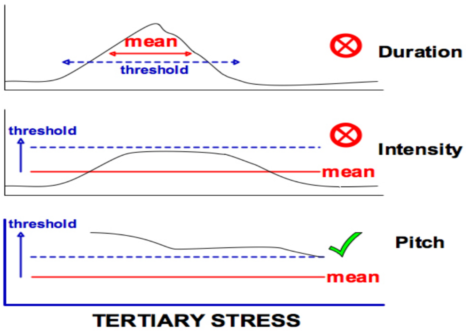


Figure 5: Rating of vowel stress levels. In this case a tertiary stress.

In the LinguaTag application, if a particular vowel crosses a threshold value defined by the user, it is promoted to a higher level of stress. By determining the overall combination of threshold values for an event, it is then defined as either a primary, secondary or tertiary stress (Figure 6):

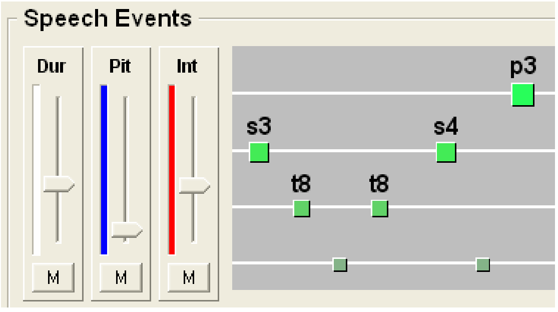


Figure 6: User specification of vowel threshold levels in LinguaTag

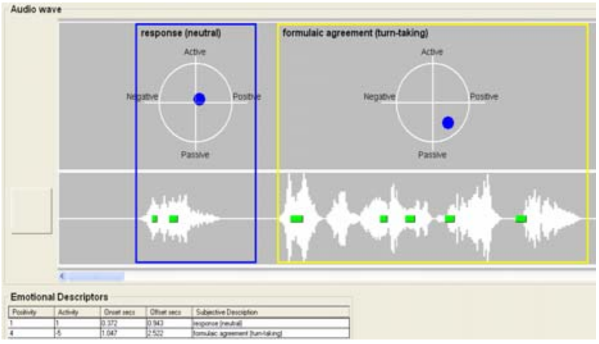
By prioritising vowel events in this manner, it is possible to determine elements of prominence in a speech signal. The analysis of such prominent events aims to provide means of focus when seeking to determine the acoustic correlates of emotional speech. The manual control of the threshold values used in stress promotion, in conjunction with user testing, can be used to develop a statistical definition of the threshold values required to accurately define stress levels. This will eventually lead to an automatic model for vowel parameter thresholds that will allow the entire stress promotion process to be performed automatically. As there are many types of information that can be extracted from a speech signal, this framework seeks to define a means of considering vowel stress information relative to its overall salience within the clip.

5.1 Querying Vowel Events for Other Acoustic Parameters

As previously mentioned, each vowel event in a speech clip is rated for stress based on its duration, intensity and pitch. Voice quality attributes such as jitter [32, 33], shimmer [34], HNR [35] and Hammarberg Index [36, 37] are also obtained for each vowel event, which can then be analysed in conjunction with duration, intensity and pitch information.

6. ANALYSIS OF EMOTIONALLY RATED SPEECH CLIPS

Determination of the acoustic correlates of emotional speech is an open research question [38, 39]. Correlations between fundamental frequency and emotional dimensions have been observed [40], but again further work is needed. The vowel stress method proposed in this paper allows prominent events to be analysed for a variety of parameters, which may prove to be acoustic correlates of emotional speech. In this process, an asset obtained using experimental mood induction procedures [5, 6] is first rated in terms of its emotional dimensions using LinguaTag, (Figure 7):



Positivity	Activity	Onset secs	Offset secs	Subjective Description
1	1	0.372	0.943	response (neutral)
4	-5	1.047	2.522	formulaic agreement (turn-taking)

Figure 7: Emotional dimension rating in LinguaTag

This user rating forms part of a statistical evaluation of the perceived emotional dimensions in a clip, whereby the collation of user group results is used to define overall values of activation and evaluation. In this manner, a statistically robust approach to the definition of emotional dimensions is made by consensus, rather than individual ratings or small expert groups [9, 41].

Once rated, the metadata relating to the acoustic parameters in each vowel event is output within a SMIL format file, which can then be queried for analysis. The querying of groups of assets within the corpus will then be used to investigate the presence of acoustic correlates within emotionally rated speech clips.

7. CONCLUSIONS

This paper describes a method of vowel stress tagging for the purposes of analysis of emotional speech assets. This analysis uses assets generated as part of an emotional speech corpus, and seeks to produce a means of determining the acoustic correlates of emotional speech. Current work is focussed on the means of processing files using the LinguaTag application, with a view to refining and improving the analysis. This refinement is also informed by statistical analysis of the assets as they are obtained, which will ideally eventually lead to a method of automatic analysis of the acoustic correlates of emotional speech.

Acknowledgments. The research leading to this paper was partially supported by the European Commission under contract FP6-027122, “Semantic AudiovisuaL Entertainment Reusable Objects- SALERO”.

8. REFERENCES

1. Cowie, R. and R.R. Cornelius, *Describing the emotional states that are expressed in speech.* Speech Communication Special Issue on Speech and Emotion, 2003. **40**(1-2): p. 5-32.
2. Campbell, N. *Databases of emotional speech*. in *ISCA Workshop on Speech and Emotion*. 2000. Northern Ireland.
3. Douglas-Cowie, E., et al., *Emotional speech: towards a new generation of databases.* Speech Communication Special Issue Speech and Emotion, 2003. **40**(1-2): p. 33–60.
4. Vaughan, B. and C. Cullen, *The Use of Task Based Mood-Induction Procedures to Generate High Quality Emotional Assets*, in *6th annual Conference on Information Technology and Telecommunications*. 2006: Carlow, Ireland.
5. Cullen, C., Vaughan, B. ,Kousidis, S., Wang, Yi ., McDonnell, C. and Campbell, D. . *Generation of High Quality Audio Natural Emotional Speech Corpus using Task Based Mood Induction* in *International Conference on Multidisciplinary Information Sciences and Technologies* 2006. Extremadura, Merida.
6. Vaughan, B., S. Kousidis, and C. Cullen, *Task-Based Mood Induction Procedures for the Elicitation of Natural Emotional Responses*, in *The 5th International Conference on Computing, Communications and Control Technologies: CCCT 2007*. 2007: Orlando, Florida, USA.
7. Scherer, K.R., *On the nature and function of emotion: A component process approach*, in *Approaches to emotion*, K.R. Scherer and P. Ekman, Editors. 1984, Erlbaum: Hillsdale, NJ. p. 293-317.
8. Cowie, R., et al., *Emotion recognition in human-computer interaction.* IEEE Signal Processing Magazine, 2001. **18**(1): p. 32-80.
9. Schroeder, M., *Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*, in *Faculty of Philosophy*. 2004, Universitat des Saarlandes. p. 288.

10. Cutler, A., D. Dahan, and W.v. Donsellar, *Prosody in the Comprehension of Spoken Language: A Literature Review.* Language and Speech, 1997(40(2)): p. 141-201.
11. Mixdorff, H. *Speech Technology, ToBI, and Making Sense of Prosody.* . in *Aix-en-Provence, France.* 2002.
12. Werner, S. and E. Keller, *Prosodic Aspects of Speech*, in *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of tha Art, and Future Challenges*, E. Keller, Editor. 1994, John Wiley: Chichester. p. 23-40.
13. Lehiste, I., *Suprasegmentals*. 1970, Cambridge, MA: MIT Press.
14. Dutoit, T., *An Introduction to Text-to-Speech Synthesis*. Text, Speech and Language Technology, ed. N. Ide and J. Veronis. Vol. 3. 1997, Dordrecht: Kluwer Academic Publishers.
15. Dauer, R.M. *Phonetic and Phonological Components of Language Rhythm*. in *11th International Congress of Phonetic Sciences*. 1987: Tallinn.
16. Ramus, F., M. Nesporb, and J. Mehlera, *Correlates of linguistic rhythm in the speech signal.* Cognition, 1999. **73**(3): p. 265-292.
17. Tsao, Y.-C., G. Weismer, and K. Iqbal, *The effect of intertalker speech rate variation on acoustic vowel space.* The Journal of the Acoustical Society of America, 2006. **119**(2): p. 1074-1082.
18. Honorof, D.N. and D.H. Whalen, *Perception of pitch location within a speaker's F0 range.* The Journal of the Acoustical Society of America, 2005. **117**(4): p. 2193-2200.
19. Farinas, J. and F. Pellegrino, *Automatic Rhythm Modeling for Language Identification*, in *Eurospeech 2001*. 2001: Scandinavia.
20. Mehler, J., et al., *Coping with linguistic diversity: The infant's viewpoint*, in *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*, J.L. Morgan and K. Demuth, Editors. 1996, Lawrence Erlbaum Associates: Mahwah, NJ. p. 101-116.
21. Ramus, F. *Acoustic correlates of linguistic rhythm: Perspectives*. in *Proceedings of Speech Prosody*. 2002. Aix-en-Provence, France.
22. Ramus, F., *Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues.*, in *Annual Review of Language Acquisition 2* 2002, John Benjamins Publishing Company.
23. Bertoncini, J., et al., *An investigation of young infants' perceptual representations of speech sounds.* Journal of Experimental Psychology: General, 1988. **117**(1): p. 21-33.
24. Pellegrino, F. and R. Andre-Obrecht. *An unsupervised approach to language identification*. in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '99*. 1999. Phoenix, AZ, USA.
25. Rouas, J.L., J.´Farinas, and F. Pellegrino. *Automatic Modelling of Rhythm and Intonation for Language Identification*. in *15th International Congresses of Phonetic Sciences, ICPhS*. 2003. Barcelona, Spain.
26. Vallee, N., et al. *Des lexiques aux syllabes des langues du monde: typologies et structures*. in *XXIII`emes Journees d’Etude sur la Parole*. 2000. Aussois, France.
27. Cullen, C., et al., *LinguaTag: An Emotional Speech Analysis Application*, in *12th World Multi-Conference on Systemics, Cybernetics and Informatics: WM-SCI '08*. 2007, Submitted Paper: Orlando, Florida, USA.
28. Boersma, P. and D. Weenink, *Praat: doing phonetics by computer*. 2006.
29. W3C (2005) *Synchronized Multimedia Integration Language (SMIL 2.1)*. **Volume**,
30. Klabbers, E. and J.P.H.v. Santen, *Clustering of foot-based pitch contours in expressive speech*, in *5th ISCA Speech Synthesis Workshop*. 2004: Pittsburgh.
31. Crystal, D., *A dictionary of linguistics and phonetics, 4th edition*. 2002, Oxford: Blackwell.
32. Ozdas, A., Shiavi, R.G., Silverman, S.E., Silverman, M.K., Wilkes, D.M., *Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk.* IEEE Biomedical Engineering,, 2004. **51**(9): p. 1530 - 1540.
33. Bernstein, J., Clayton, K, J., Dobrian, C., DuBois, L, R., Gerstmann, D., Jones, R., Nevile, B., and and G. Taylor, *Jitter Tutorial*. 2006, Cycling'74. p. 541.
34. Gobl, C., E. Bennett, and A.N. Chasaide. *Expressive Synthesis: How Crucial is Voice Quality?* in *IEEE Workshop on Speech Synthesis*. 2002. Santa Monica, CA (USA).
35. Severin, F., B. Bozkurt, and T. Dutoit. *HNR extraction in voiced speech, oriented towards voice quality analysis*. in *European Signal Processing Conference, EUSIPCO'05*. 2005. Antalya,Turkey.
36. Fant, G. and Q. Lin, *Comments on glottal flow modelling and analysis*, in *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms*, J. Gauffin and B. Hammarberg, Editors. 1991, Singular Publishing Group: San Diego. p. 47-56.
37. Lee, C.K. and D.G. Childers, *Some acoustical, perceptual, and physiological aspects of vocal quality*, in *Vocal Fold Physiology: Acoustic, Perceptual, and Physiological Aspects of Voice Mechanisms*, J. Gauffin and B. Hammarberg, Editors. 1991, Singular Publishing Group: San Diego. p. 233-242.
38. Banse, R. and K.R. Scherer, *Acoustic profiles in vocal emotion expression.* Journal of Personality and Social Psychology, 1996. **70**(3): p. 614-636.
39. Schroeder, M., *Emotional Speech Synthesis: A Review.* Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH'01), 2001. **1**: p. 561-564.
40. Schröder, M., et al., *Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis.* Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH'01), 2001. **1**: p. 87-90.
41. Douglas-Cowie, E., R. Cowie, and M. Schröder. *A new emotion database: considerations, sources and scope*. in *ISCA Workshop on Speech and Emotion*. 2000. Northern Ireland.