2008-01-01

LinguaTag: an Emotional Speech Analysis Application

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

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

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

LinguaTag: an emotional speech analysis application

Abstract: *The analysis of speech, particularly for emotional content, is an open area of current research. Ongoing work has developed an emotional speech corpus for analysis, and defined a vowel stress method by which this analysis may be performed. This paper documents the development of LinguaTag, an open source speech analysis software application which implements this vowel stress emotional speech analysis method developed as part of research into the acoustic and linguistic correlates of emotional speech. The analysis output is contained within a file format combining SMIL and SSML markup tags, to facilitate search and retrieval methods within an emotional speech corpus database. In this manner, analysis performed using LinguaTag aims to combine acoustic, emotional and linguistic descriptors in a single metadata framework.*

1. Introduction

Existing work in the field of emotional speech research has foccused on the means by which suitable speech assets may be obtained [1-3], leading to work towards the creation of a corpus of natural emotional speech [4-6]. Analysis of such assets can then be performed to determine the potential acoustic correlates of emotional speech, and to this end a speech analysis application called Linguatag was developed. LinguaTag can be used to produce tagged speech audio files for analysis [7] and retrieval. Linguatag aims to provide means of querying various acoustic, linguistic and emotional parameters of speech audio files for the purposes of analysis and retrieval.

2. Application Design

The LinguaTag application is written in Eiffel, and runs Praat [8] source scripts for audio analysis of speech signals in the WAV format (Figure 2). The results of these Praat queries are then displayed in a GUI for editing and analysis before being output in a standard SMIL file format.

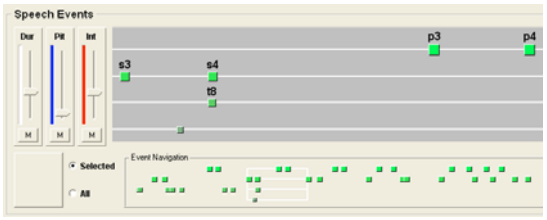

Figure 1: Screenshot of the LinguaTag emotional speech analysis application

The inclusion of manual linguistic and emotional dimension analysis within the application allows the user to define various important aspects of a speech signal within a single common file format, which can then be used in processes such as lip-synching animation and content storage and retrieval (Figure 1):

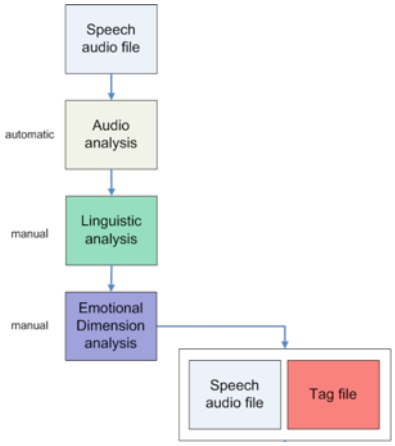

Figure 2: Workflow diagram of the LinguaTag application

LinguaTag uses the Praat analysis engine 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 [9].

3. Vowel stress analysis

The isolation of vowel events in a speech signal is a common approach in speech analysis [10-13]. Although a syllable may be formed around non-vocalic events, most speech patterns involve the alternation of vowels and consonants [14]. The definition of the 'pseudo-syllable' [15] is based on the observation that the CV structure is the most common structure [13, 16], and thus leads to the use of a vowel onset detection algorhythm to determine the occurrence of each vowel (and hence each CV) in a speech event. Prosody in speech can be considered in many different ways [17, 18], 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 [19, 20]. To facilitate comparison, Dutoit [21] 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 [21]

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 [22], particularly as it relates to acoustic parameters [21]. The acoustic attributes of pitch, intensity and duration were chosen as fundamental features of a speech event, which are agreed as being common to all 3 representational models of speech [21]. Using this method, a vowel event can thus be graded in terms of threshold values relative to these three parameters:

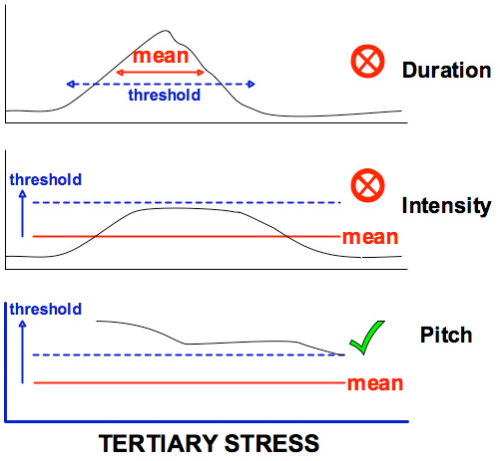

Figure 3: Rating of vowel stress levels

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 4):

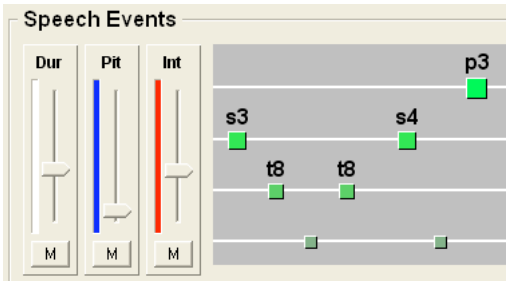

Figure 4: User specification of vowel threshold levels in LinguaTag

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. There are many types of information that can be extracted from a speech signal, and this framework seeks to define a means of considering this information relative to its overall salience within the clip.

4. Analysis of emotional speech clips

4.1. Acoustic Analysis

Vowel events in the speech signal are queried for pitch and intensity information, and this information is displayed on-screen using simple contour definitions. These contour descriptors are adapted from existing research into data sonification [23-25], which is based on the work of ethnomusicologist Charles Adams [26] into the classification of Native American melodies. Adams suggested that a contour could be defined in terms of 4 minimal boundary pitches: initial pitch (I), highest pitch (H), lowest pitch (L) and final pitch (F). The relations between these 4 boundary pitches were then summarised in 3 categories:

Slope, S- slope defines a comparison between the initial (I) and final (F) pitches as either ascending, level or descending.

Deviation, D- changes in direction between boundary pitches specify levels of deviation. Thus if all four pitches are equal then the deviation is zero, with subsequent disparities between any of the 4 giving different levels of deviation.

Reciprocal, R- The direction of the first deviation (either I to H or I to L) is referred to as its reciprocal, dictating the direction of the overall contour.

Using these features as a template, Adams defined 15 basic contour shapes for melodic classification (Figure 5).

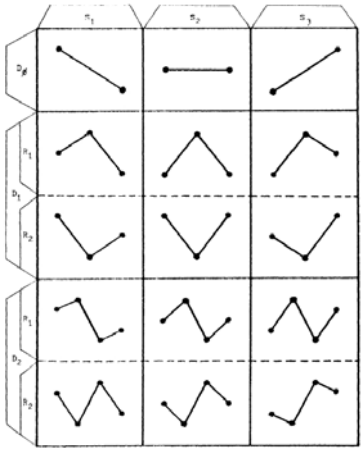

Figure 5: 15 Contour Graphs, taken from Adams [26]

This approach performed well for defining melodies that had been reduced to groups of 4 salient pitches, and allowed Adams to define the similarities and differences between music from 2 separate Native American tribes.

By using a similar boundary pitch classification method in speech analysis within LinguaTag, it is possible for users to define speech contour characteristics quickly from this template set (Figure 6):

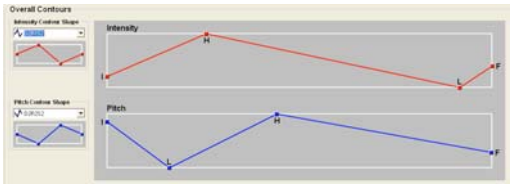

Figure 6:LinguaTag contour analysis screen

Once rated, the defined contours can then be queried as metadata in relation to other aspects of the speech signal, such as its emotional dimensions (see section 5).

4.2. Voice Quality Analysis

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 [27, 28], shimmer [29], HNR [30] and Hammarberg Index [31, 32] are also obtained for each vowel event, which can then be analysed in conjunction with duration, intensity and pitch information. All analysis information is displayed within the application, allowing the user to manually check for specific voice quality values (Figure 7):

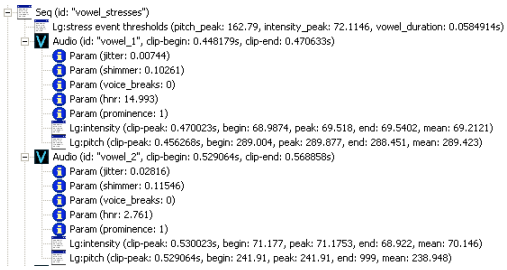

Figure 7: LinguaTag output file display screen

4.3. Linguistic Analysis

The LinguaTag application allows the speech audio file to be edited to demark areas of linguistic interest (Figure 8):

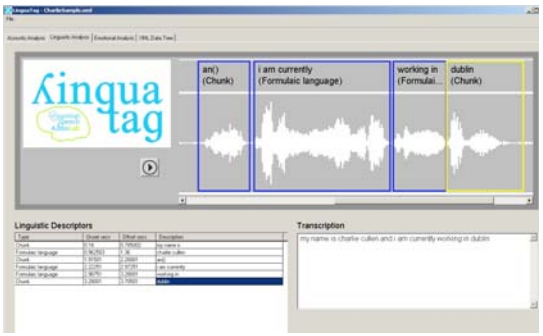

Figure 8: Linguistic analysis screen

Linguistic features such as formulaic language and chunks can be annotated in the current version, and it is intended to include provision for many other salient features of linguistic relevance such as speed of delivery, power relationship and prominence in later versions. These features will allow the LinguaTag application to be used for more effective and multidisciplinary analysis of a speech signal in a single pass.

5. Emotional Analysis

In this research, circumplex emotional modelling [33-35] is used to rate speech assets on unit scales of activation and evaluation (Figure 9):

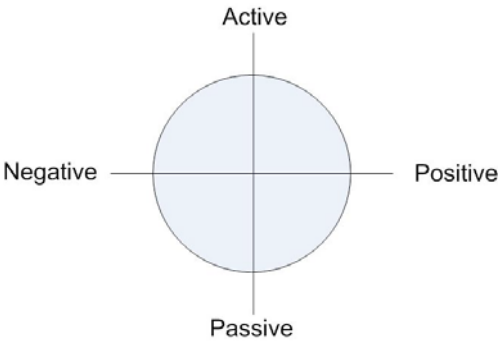

Figure 9: Circumplex emotional model denoting dimensions of activation and evaluation, adapted from Scherer [33]

This dimensional rating allows the speech corpus to be defined in terms of its emotional content (using statistical listening tests [5]) prior to acoustic analysis using LinguaTag.

Determination of the acoustic correlates of emotional speech is an open research question, with no definitive results being available at the present time [36, 37]. Correlations between fundamental frequency and emotional dimensions have been observed [38], but again further work is needed. LinguaTag 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 manually rated in terms of its emotional dimensions (Figure 10):

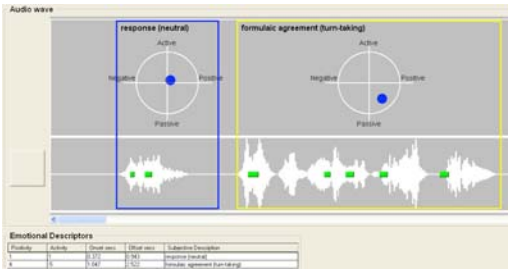

Figure 10: Emotional dimension rating in LinguaTag

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 [35, 39].

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

6. Conclusions

LinguaTag is an application developed as part of ongoing research into the production, analysis and retrieval of emotional speech clips. The application implements the vowel stress emotional speech analysis method used by this research, and provides and effective user interface for the manual annotation of automatically analysed speech assets. Work on LinguaTag is ongoing, with future developments seeking to standardise output files further to conform with the IMDI speech corpus metadata standard [40]. It is hoped that this will move some way towards defining the acoustic and linguistic correlates of emotional speech.

7. Acknowledgements

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] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication Special Issue on Speech and Emotion*, vol. 40, pp. 5-32, 2003.

[2] N. Campbell, "Databases of emotional speech," presented at ISCA Workshop on Speech and Emotion, Northern Ireland, 2000.

[3] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: towards a new generation of databases," *Speech Communication Special Issue Speech and Emotion*, vol. 40, pp. 33–60, 2003.

[4] B. Vaughan 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*. Carlow, Ireland, 2006.

[5] C. Cullen, 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 " presented at International Conference on Multidisciplinary Information Sciences and Technologies Extremadura, Merida, 2006.

[6] B. Vaughan, 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*. Orlando, Florida, USA, 2007.

[7] C. Cullen, B. Vaughan, S. Kousidis, and F. Reilly, "A vowel-stress emotional speech analysis method," in *Knowledge Acquisition from Multimedia Content, KAMC 2007*. Genoa, Italy: Submitted Paper, 2007.

[8] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," 4.4.18 ed, 2006.

[9] W3C, "Synchronized Multimedia Integration Language (SMIL 2.1)," 2005.

[10] F. Ramus, M. Nesporb, and J. Mehlera, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, pp. 265-292, 1999.

[11] Y.-C. Tsao, G. Weismer, and K. Iqbal, "The effect of intertalker speech rate variation on acoustic vowel space," *The Journal of the Acoustical Society of America*, vol. 119, pp. 1074-1082, 2006.

[12] D. N. Honorof and D. H. Whalen, "Perception of pitch location within a speaker's F0 range," *The Journal of the Acoustical Society of America*, vol. 117, pp. 2193-2200, 2005.

[13] J. Farinas and F. Pellegrino, "Automatic Rhythm Modeling for Language Identification," in *Eurospeech 2001*. Scandinavia, 2001.

[14] F. Pellegrino and R. Andre-Obrecht, " An unsupervised approach to language identification," presented at Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '99, Phoenix, AZ, USA, 1999.

[15] J. L. Rouas, J. ´Farinas, and F. Pellegrino, "Automatic Modelling of Rhythm and Intonation for Language Identification," presented at 15th International Congresses of Phonetic Sciences, ICPhS, Barcelona, Spain, 2003.

[16] N. Vallee, L. J. Boe, I. Maddieson, and I. Rousset, "Des lexiques aux syllabes des langues du monde: typologies et structures," presented at XXIII`emes Journees d'Etude sur la Parole, Aussois, France, 2000.

[17] A. Cutler, D. Dahan, and W. v. Donsellar, "Prosody in the Comprehension of Spoken Language: A Literature Review," *Language and Speech*, pp. 141-201, 1997.

[18] H. Mixdorff, "Speech Technology, ToBI, and Making Sense of Prosody. ," presented at Aix-en-Provence, France., 2002.

[19] S. Werner 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, Ed. Chichester: John Wiley, 1994, pp. 23-40.

[20] I. Lehiste, *Suprasegmentals*. Cambridge, MA: MIT Press, 1970.

[21] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, vol. 3. Dordrecht: Kluwer Academic Publishers, 1997.

[22] R. M. Dauer, "Phonetic and Phonological Components of Language Rhythm," presented at 11th

International Congress of Phonetic Sciences, 1987.

[23] C. Cullen and E. Coyle, "Harmonic Combination of Contour Icon Patterns," in *Irish Signals and Systems Conference (ISSC)*. Dublin, Ireland, 2006.

[24] C. Cullen and E. Coyle, "Musical Pattern Design Using Contour Icons," in *International Computer Music Conference (ICMC)*. Louisiana, USA, 2006.

[25] C. Cullen and E. Coyle, "Information Delivery on Mobile Devices Using Contour Icon Sonification," in *Irish Signals and Systems Conference, ISSC*. Dublin, Ireland, 2005.

[26] C. Adams, "Melodic Contour Types," *Ethnomusicology*, pp. 179-215, 1976.

[27] A. Ozdas, 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,*, vol. 51, pp. 1530 - 1540, 2004.

[28] J. Bernstein, Clayton, K, J., Dobrian, C., DuBois, L, R., Gerstmann, D., Jones, R., Nevile, B., and and G. Taylor, "Jitter Tutorial," Cycling'74, 2006, pp. 541.

[29] C. Gobl, E. Bennett, and A. N. Chasaide, "Expressive Synthesis: How Crucial is Voice Quality?," presented at IEEE Workshop on Speech Synthesis, Santa Monica, CA (USA), 2002.

[30] F. Severin, B. Bozkurt, and T. Dutoit, "HNR extraction in voiced speech, oriented towards voice quality analysis," presented at European Signal Processing Conference, EUSIPCO'05, Antalya,Turkey, 2005.

[31] G. Fant 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, Eds. San Diego: Singular Publishing Group, 1991, pp. 47-56.

[32] C. K. Lee 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, Eds. San Diego: Singular Publishing Group, 1991, pp. 233-242.

[33] K. R. Scherer, "On the nature and function of emotion: A component process

approach," in *Approaches to emotion*, K. R. Scherer and P. Ekman, Eds. Hillsdale, NJ: Erlbaum, 1984, pp. 293-317.

[34] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, pp. 32-80, 2001.

[35] M. Schroeder, "Speech and Emotion Research. An Overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis," in *Faculty of Philosophy*: Universitat des Saarlandes, 2004, pp. 288.

[36] R. Banse and K. R. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology*, vol. 70, pp. 614-636, 1996.

[37] M. Schroeder, "Emotional Speech Synthesis: A Review," *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH'01)*, vol. 1, pp. 561-564, 2001.

[38] M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen, "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis," *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH'01)*, vol. 1, pp. 87-90, 2001.

[39] E. Douglas-Cowie, R. Cowie, and M. Schröder, "A new emotion database: considerations, sources and scope," presented at ISCA Workshop on Speech and Emotion, Northern Ireland, 2000.

[40] ISLE, "IMDI (ISLE Metadata Initiative), Metadata Elements for Session Descriptions," Draft Proposal Version 3.0.3 ed, 2003.