Intonation Template Matching for Syllable-Level Prosody Encoding

Abdul Rehman[1], Jian Jun Zhang[1] and Xiaosong Yang[1]

[1]*Faculty of Media and Communications, Bournemouth University, Bournemouth, United Kingdom*

Abstract

We address the challenge of machine interpretation of subtle speech intonations that convey complex meanings. We assume that emotions and interrogative statements follow regular prosodic patterns, allowing us to create an unsupervised intonation template dictionary. These templates can then serve as encoding mechanisms for higher-level labels. We use piecewise interpolation of syllable-level formant features to create intonation templates and evaluate their effectiveness on three speech emotion recognition datasets and declarative-interrogative utterances. The results indicate that individual syllables can be detected for basic emotions with nearly double the accuracy of chance. Additionally, certain intonation templates exhibit a correlation with interrogative implications.

Keywords

intonations, speech processing, emotion recognition, computational paralinguistics,

1. Introduction

Paralingual intonations can alter sentence meanings without changing the words, such as adding sarcasm or conveying politeness. These nuances are often challenging for machine speech recognition due to the ambiguity of implications [1]. Nonetheless, some implied meanings in speech are thought to be consistent across cultures and languages [2].

This research aims to simplify the mel-spectrum into a few explainable variables that capture paralingual cues in temporal sequences. Traditional sequence learning methods like RNN have limitations due to dataset variations and domain-specific issues [3, 4]. To address this, we propose using the quantifiable flux of syllables rather than the discrete sequence of phonemes. We create standard templates from common syllable feature patterns and use them to match test syllables, offering a way to mathematically quantify syllable flux without relying on a standard paralingual symbol dictionary.

There have been few pieces of research done on the comparison between the applicability of prosodic and lexical features for emotional clue extraction, they point out the higher importance of prosodic features as compared to their lexical dependence [5, 6]. Moreover, the prosodic cues

Cognitive AI 2023, 13th-15th November, 2023, Bari, Italy.

arehman@bournemouth.ac.uk (A. Rehman); jzhang@bournemouth.ac.uk (J. Zhang); xyang@bournemouth.ac.uk (X. Yang)

https://staffprofiles.bournemouth.ac.uk/display/arehman (A. Rehman); https://jzhang.bournemouth.ac.uk/ (J. Zhang); https://staffprofiles.bournemouth.ac.uk/display/xyang (X. Yang)

0000-0003-2345-2256 (A. Rehman); 0000-0002-7069-5771 (J. Zhang); 0000-0003-3815-0584 (X. Yang)

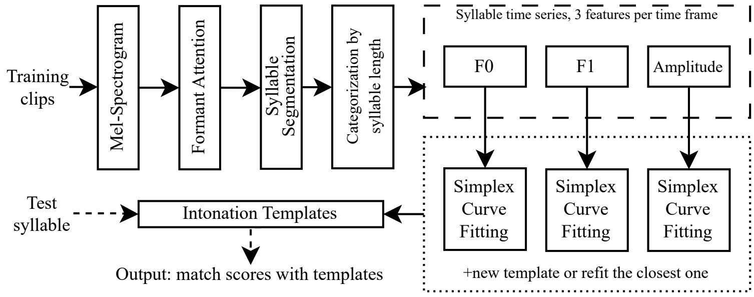

Figure 1: The overview of the proposed method to encode intonations.

are considered ambiguous and speaker-dependent for the human perception as well, i.e., human listeners also adjust their prosody perception based on priors [7].

Intonation detection is relatively a less researched area due to the lack of proper datasets. There are a few studies that use the lexical and acoustic features to recognize interrogative intonation [3, 8, 9, 10]. They all report better detection using lexicon instead of acoustics. Margolis et al. reported that their text-trained model had poor recognition of declarative questions even when considering the prosodic features [8]. The poor recognition is thought to be due to the ambiguous implications of uncertainty, i.e., the rising pitch at the end doesn't always mean interrogation, it can also signal confirmation seeking or uncertainty [11].

2. Method

The proposed method is based on the syllable formant attention mechanism, which relies on the first two formants of vowel sounds as effective descriptors of speech [12, 13, 14]. We also incorporate the combined amplitude of the top 3 formants as a third frame-level feature, assuming that loudness variation conveys intonation information. Figure 3 illustrates these formants in a 4-syllable speech segment, segmented using onset and offset detection as previously described [12]. To quantify these formant patterns as a measure of intonation, we create a template dictionary from common formant patterns using syllable feature-time data interpolation. We then assess how well other syllables align with these templates.

The final output of the interpolation process is the residual errors and correlations with the feature templates that can be used to estimate useful paralingual cues. For a syllable template s of a feature h (out of $F0, F1, Amp$) there are five polynomials $\beta_{s,h}$ in a 4-degree polynomial equation that creates a template model:

$$y_{s,h}(t_i) = \beta_{s,h,0} + \beta_{s,h,1}t_i + \beta_{s,h,2}t_i^2 + \beta_{s,h,3}t_i^3 + \beta_{s,h,4}t_i^4 \tag{1}$$

where i is the index of the syllable being fitted, $y_{s,h}$ is the interpolated value of the formant feature h at frame index t_i, and s is the incremental index of the equation in the feature template dictionary. We want $y_{s,h}$ to be as close to the actual value of that feature as possible. A loss minimization method is used that fits the know formant features onto a polynomial curve in two stages. The first stage creates an estimate of best-fit for the polynomial matrix $\beta_{s,h}$ using matrix manipulations as

$$\beta_{s,h} = [X \cdot X^T]^{-1} \cdot [Y^T \cdot X^T] \tag{2}$$

where X is a matrix with frames index t of all the syllables to be fitted as rows and the polynomial coefficients ($p \in \{0, ..., 4\}$ for quartic fitting) as columns where each element has a value of $X_{t,p} = t^p$. Whereas Y is the single-column matrix of the expected formant feature values at each frame of the syllable. Then at the second stage, a loss minimization by gradient descent is used to improve the fitting of the polynomials. We used the simplex algorithm for unconstrained minimization of the polynomial regression loss [15].

A sum-squared error is used to estimate $loss_{s,h}$ during the unconstrained minimization as

$$loss_{s,h} = \sum_{i=0}^{N_s} \sum_{t_i=0}^{len_i} [y_{s,h}(t_i) - z_{s,h}(t_i)]^2 \tag{3}$$

where $y_{s,h}$ is the calculated values by the polynomial equation, and $z_{s,h}$ is the actual value of the formant feature h of syllable s at frame t_i, N_s is the total number of syllables in being fitted for this template, and len_i is the total frame length of an individual syllable i. The initial coefficients $\beta_{s,h}$ produced in the first stage of coefficients are only a rough estimate to help reach the minima in less time. The actual unconstrained minimization of the polynomial curve is performed using the simplex algorithm. The algorithm moves towards the function $loss_{s,h}$ minimum by adjusting the parameters of the worst points. The simplex algorithm usually converges in 5 to 20 iterations because of the initialization, otherwise, it takes many more iterations. Figure 3 shows the regression values of quartic curves of the first formants of each syllable laid over the actual frequencies.

2.1. Matching Syllables with Intonation Template

The sequences of the first two formants; $F0$ and $F1$, and the magnitude Amp are fitted using audio speech recordings to create 3 separate sets of curve templates using the curve fitting method described above. The 3 sets are further divided into 10 (for each of the 3 features) categories by length because we assume that the duration of a syllable is also one of the key discriminating features. Feature sequences of various durations have their own sets of template curves. The number of total templates for a feature-duration class depends on the variety in the recording and on the coefficient of determination R^2 threshold set for a match to be considered or to create a new template if the match score is lower than the threshold.

Once all the template coefficients have been estimated, they can be used as a match predictor model for various paralingual tasks such as emotion recognition or interrogative speech detection. For example, if there are a total of 100 templates of various shapes and sizes for 3 formant features, the R^2 scores for all of the 100 templates can be used as a feature vector to train a classifier to predict any paralingual label for a syllable.

3. Experimentation

We evaluated the proposed method on the tasks of speech emotion recognition and interrogative intonation analysis. We make 2 assumptions to evaluate our proposed approach:

- Basic emotions can be recognized from individual syllables without needing a huge dataset for regularization. We test this in two ways: 1) By training speech emotion recognition models on one dataset and testing on another, and 2) by decreasing the size of data used for training to check if the proposed method can be trained with a small sample.
- Interrogative intonation can be distinguished from individual syllables when the same statement is said with a declarative vs interrogative tone. Due to the lack of a dataset that can be used for machine learning, we used a small dataset that was only big enough for observations.

For speech emotion recognition, we used 3 widely used databases that are recorded in a scripted or improvised scenario in English. IEMOCAP database [16], MSP-Improv database [17] and the RAVDESS speech dataset [18]. The validation was performed using a 5-fold method when the training source data and testing data are from the same database. Whereas, for cross-corpus validation, two different datasets were used for training and testing.

For interrogative intonation analysis, we used a small set of 72 utterances that were originally collected for the purpose of controlled stimuli by Xie et al. [7]. There are 6 two-word sentences ending with a continuous verb "ing" sound, said 12 times with an increasingly questioning tone by the same speaker on a continuum between statements and declarative questions e.g., "It's raining." and "It's raining?". The lack of variability in other factors makes this dataset ideal for an analysis of the effectiveness of the proposed template-matching method.

Using the method described in Section 2, a set of feature templates was derived from the training dataset. The templates' polynomial coefficients are used to estimate correlations R^2 of each formant feature ($F0, F1, Amp$) with their respective sets of templates. The template that matches the test syllable's feature sequence the most would have the highest correlation among all the template curves. The number of template curves is not fixed and depends on the error threshold set during training. In relative terms, the error threshold was set to 0.08, i.e., if the nearest match has $R^2 \leq 0.92$ for a new syllable with the already known templates then a new template is created using the new syllable. Otherwise, the new syllable data is appended to the data series of the nearest matching template, which is then refitted again using the simplex algorithm.

The encoded features vectors extracted from the proposed method (the match scores with the templates) were then used to train a single hidden layer MLP (Multi-Layer Perceptron) with 8 units to perform the emotion classification task. For the cross-corpus validation tasks, the width of the syllable feature vector for training on RAVDESS was 133, 150 for IEMOCAP, and 146 for MSP-Improv.

3.1. Results

The results in Table 1 show that the UA (Unweighted Accuracy) for the proposed method is significantly better than the chance (25%) that reflects the amount of prosodic information

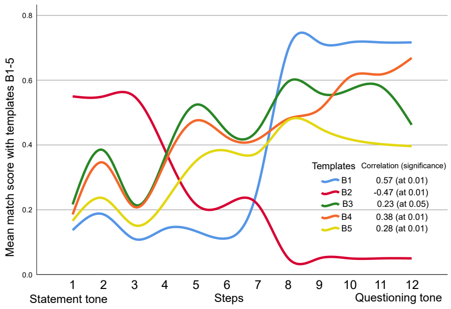

Figure 2: Mean match scores of templates for 6 two-word sentences spoken 12 times for each step on a declarative-interrogative intonation continuum.

captured by templates. More importantly, the cross-corpus accuracies show that templates learned from one corpus can predict the emotions in other corpora. For comparison, Alex et al. [19] have reported syllable-level accuracies for the IEMOCAP dataset 37.57% and 63.83% at utterance level. Most other works report accuracies for utterance level therefore a valid comparison can't be made [20].

Table 1
Syllable-level UA% for emotion classification (4 classes) using cross-corpus or 5-fold validation.

	Target		
Source	RAVDESS	IEMOCAP	MSP-Improv
RAVDESS	52.3	47.8	38.1
IEMOCAP	47.9	49.6	41.4
MSP-Improv	40.3	46.5	44.2

Figure 2 shows the plots of means of the match scores with five syllable-feature templates. Only the most correlated 5 of the total 28 generated templates are shown. There is a significant correlation between the top 2 templates with the question tone continuum.

4. Conclusions

We introduced a method for analyzing syllable-level intonation patterns by matching formant features in syllables with common patterns. Using template similarity scores, we predicted the emotional tone of each syllable and explored the correlation between match scores and questioning intonation levels. Our findings revealed that only a few templates effectively captured interrogative intonation in the sample recordings.

This research had limitations, including the absence of syllable-level annotated data, which affected learning precision due to variations within utterances. Future work involves collecting and annotating syllable-level data to enhance computational paralinguistics. Another challenge was the computational heaviness of polynomial optimization for large datasets. Future efforts will explore alternative approaches for syllable template extraction to improve efficiency.

References

[1] S. G. Kanakaraddi, S. S. Nandyal, Survey on parts of speech tagger techniques, in: 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), IEEE, 2018, pp. 1–6.

[2] M. E. Armstrong, M. del Mar Vanrell, Intonational polar question markers and implicature in american english and majorcan catalan, Speech Prosody 2016 (2016) 158–162.

[3] Y. Tang, Y. Huang, Z. Wu, H. Meng, M. Xu, L. Cai, Question detection from acoustic features using recurrent neural network with gated recurrent unit, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 6125–6129.

[4] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, M. Hao, Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence, Information Sciences 563 (2021) 309–325.

[5] I. Grichkovtsova, A. Lacheret, M. Morel, The role of intonation and voice quality in the affective speech perception, in: Eighth Annual Conference of the International Speech Communication Association, 2007.

[6] I. Grichkovtsova, M. Morel, A. Lacheret, The role of voice quality and prosodic contour in affective speech perception, Speech Communication 54 (2012) 414–429.

[7] X. Xie, A. Buxó-Lugo, C. Kurumada, Encoding and decoding of meaning through structured variability in intonational speech prosody, Cognition 211 (2021) 104619.

[8] A. Margolis, M. Ostendorf, Question detection in spoken conversations using textual conversations, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011, pp. 118–124.

[9] K. Boakye, B. Favre, D. Hakkani-Tür, Any questions? automatic question detection in meetings, in: 2009 IEEE Workshop on Automatic Speech Recognition & Understanding, IEEE, 2009, pp. 485–489.

[10] A. Ando, R. Asakawa, R. Masumura, H. Kamiyama, S. Kobashikawa, Y. Aono, Automatic question detection from acoustic and phonetic features using feature-wise pre-training., in: INTERSPEECH, 2018, pp. 1731–1735.

[11] M. Šafářová, The semantics of rising intonation in interrogatives and declaratives, in: Proceedings of sinn und bedeutung, volume 9, 2005, pp. 355–369.

[12] A. Rehman, Z.-T. Liu, M. Wu, W.-H. Cao, C.-S. Jiang, Speech emotion recognition based on syllable-level feature extraction, Applied Acoustics 211 (2023) 109444.

[13] A. Rehman, Z.-T. Liu, J.-M. Xu, Syllable level speech emotion recognition based on formant attention, in: CAAI International Conference on Artificial Intelligence, Springer, 2021, pp. 261–272.

[14] R. D. Kent, H. K. Vorperian, Static measurements of vowel formant frequencies and bandwidths: A review, Journal of communication disorders 74 (2018) 74–97.

[15] J. E. Dennis Jr, D. J. Woods, Optimization on microcomputers. the nelder-mead simplex algorithm, Technical Report, 1985.

[16] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, S. S. Narayanan, Iemocap: Interactive emotional dyadic motion capture database, Language resources and evaluation 42 (2008) 335.

[17] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, E. M. Provost, Msp-improv: An acted corpus of dyadic interactions to study emotion perception, IEEE Transactions on Affective Computing 8 (2016) 67–80.

[18] S. R. Livingstone, F. A. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, PloS one 13 (2018).

[19] S. B. Alex, L. Mary, B. P. Babu, Attention and feature selection for automatic speech emotion recognition using utterance and syllable-level prosodic features, Circuits, Systems, and Signal Processing 39 (2020) 5681–5709.

[20] Z. Aldeneh, E. M. Provost, Using regional saliency for speech emotion recognition, in: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2017, pp. 2741–2745.

A. Figures

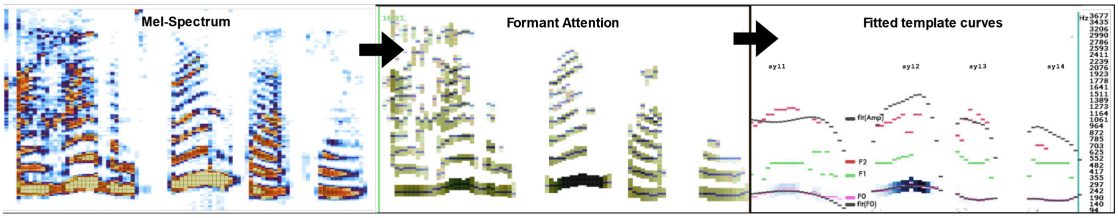

Figure 3: The spectrum plot of 3 features; F0 (fundamental frequency), F1 (second formant's frequency) and Amp (amplitude) for a sample syllable of 33 frames length (*t*), and the fitted curves of best matches found in the template dictionary for F0 and Amp (both template curves in grey, Amp not to scale. Template curves for F1 is not shown to avoid clutter).

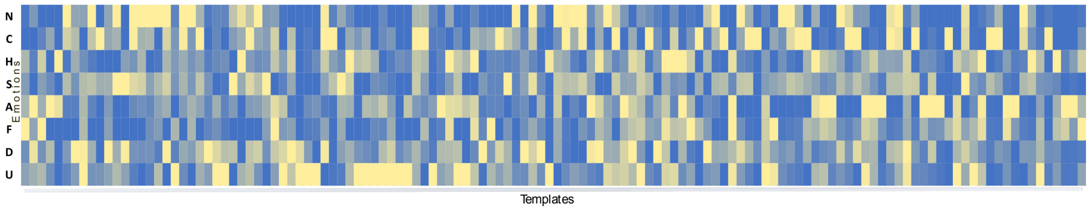

Figure 4: This raster illustrates the number of matches (yellow is higher) for 133 intonation templates in 8 emotional classes given in the RAVDESS speech dataset. The templates are derived from the same dataset.