INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH AND ANALYSIS

ISSN(print): 2643-9840, ISSN(online): 2643-9875

Volume 09 Issue 01 January 2026

DOI: 10.47191/ijmra/v9-i1-02, Impact Factor: 8.266

Page No. 4-9

Evaluation and Modeling of Spoken Expression for Synthesis

Raghdah Adnan Abdulrazzq

College Administration and Economics, University of Kirkuk, Iraq

ABSTRACT: One of the most pressing issues in modern studies of human-computer interaction is emotion-aware computing; a promising area of study that promises to provide major advances in the near future is genuine expressive speech synthesis. In this study, we detail our continuing efforts to construct expressive text-to-speech synthesis systems by creating a data-driven framework for annotation, modeling, and analysis of expressive speech. Here we detail the data-driven approach, which includes features like expression grouping and aural analysis as well as a web-based platform for voice recognition and annotation. There are also some promising signs for future study in the form of preliminary findings.

KEYWORDS: human-computer, emotion computing, synthesis system, web based

I. INTRODUCTION

Human-computer speech systems are getting more and more study interest because computers and internet services are so common in our daily lives. Conversation systems of the future and Natural speech interfaces will likely be based on expression, affect, and emotion. This means that how computers understand and expressive speech patterns and create emotional is becoming a very important problem [1,2].

aware spoken- emotion engagement looks at the emotional context of speech rather than the linguistic context. Usually, speech recognition and synthesis study looks at the linguistic content. It focuses on how something is said instead of what it is said, which is very important when dealing with the wide rangemeanings that make up speech and tones.

In order to address some of the most challenging issues in language and speech processing, researchers from several disciplines work together in the field of expressive voice synthesis [3, 4]. Many fundamental concerns about the function and storage of the abundant paralinguistic information in human speech remain unresolved [5]. These concerns may be explored via several domains, including social science, cognitive science, psychology, and natural language processing, among others.

At the moment, most methods for creating expressive speech synthesis either use simple changes to try to find or prosodic quantities and model universalities in how expression shows up in different speakers. As a result, the results are not very good. If you use a fixed set of feelings, you miss out on a lot of interesting problems and uses because these models can't catch the wide range of expressions and subtleties of human speech. Most of the time, you can't expect full-blown feelings because it's not possible to show yourself in a lot of different ways using such a simple set of emotions. In many speech situations, other emotional speaking styles are used along with or instead of showing feeling. That's why storytelling apps and audiobooks in general are so interesting: the goal is to make listening to them just as interesting as listening to a real person tell the story. It's one of many situations where having a large collection of expressive patterns can be useful. This is based on study about how expression is perceived.

We want to use computers to study the parts and traits of human speech that make it expressive so that emotionally-aware speech synthesis systems can be made. We also want to use these computers to study, create patterns of expressive speech by using a corpus-based and data-driven methodology. A trainable data-driven approach, as opposed to the more conventional model-based category analysis, would be superior because it can detect and patterns and latent structure and restore unpleased regularities in the prosodic and auditory components of expressive human speech [6,7]. These findings are consistent with the current understanding of text-to-speech (TTS) synthesis. TTS based on Random forest and Unit-Selection (US) are two corpus-based approaches that have shown the greatest promise. Perceptions of the individual speaker and expressive speech, the recorded corpus are all better addressed by this approach. By incorporating it correctly with the US or RANDOM FOREST method, it would include the whole scope of expressive speech analysis, modeling, and combination [6]. In other words, this

approach seeks to efficiently, trainably, and integratively identify the speaker's fundamental structure and emotional speaking styles that are directly related to the speaker and the speech corpus.

This article presents the results of continuing attempts to develop a data-driven, general-purpose machine learning framework for expressive speech modeling, labeling and analysis. The final objective is to develop expressive text-to-speech systems that are compatible with human language comprehension.

Fig 1. A framework for studying, imitating, and making expressive speech and text

Furthermore, the method used here can be used in other areas and modes where natural collections of "recorded" true data are needed and are available, as long as the right changes are made. Face emotions, writing with meta-annotation, and other things are examples of these kinds of media. In general, the driven - data approach is connected to the expansive area of cognitive information communications due to the fact that it is founded on data and the manner in which it is able to capture any cognitive processes that entail synergy across many modalities.

This can then be used to make communication and interaction between humans and machines more natural and emotional [8,9].

Here's how the rest of the paper is put together. The synthesis approach and expressive speech analysis is explained in Section II, which lists the steps that need to be taken and talks about some important ideas. In Section III, we talk about the online speech recognition and annotation system that was built. In Section IV, we show how to get at hidden emotional traits and the first results of the grouping methods. Section V, "Conclusion and Further Work," is the last part of the study [10,11].

II. FRAMWORK FOR SYNTHESIS AND EXPRESSIVE SPEECH ANALYSIS

All three phases—synthesis and modeling and speech analysis—are depicted in Figure 1. To begin, the phoneme-level alignment of the original script (text) and the voice recording is checked against a speech corpus that contains various forms of speech (e.g., tales, audiobooks, news broadcasts, etc.). The phoneme or word level processing of the speech is facilitated by this. In order to gain a better understanding of the language's surface structure, such as the words' part-of-speech, the presence of pauses in their context, higher-order syntactic constructs, and more, a number of natural language processing tools can be employed prior to the acoustic analysis. These features can be binary or categorical in nature. Then, the acoustic qualities may be enhanced using these features. When processing audio, several popular temporal and spectral analysis methods are used to extract a collection of prosodic and acoustic characteristics from expressive voice corpora at various granularities (sentence, word, phrase, or frame). Pitch, loudness, speaking pace, timbre, intensity, spectral band energy distributions, and other speech features are often used in speech analysis because they are believed to be significant for determining expression and mood. Their differentials of the first and second order are often considered as well [12,13]. Prosodic characteristics, spectral features excitation source features and vocal tract features are the three main components of speech that audio analysis typically aims to isolate and combine [14]. Based on the preliminary findings in [15], we use the set of characteristics discussed in Section IV of this study. Obviously, not everything is on this list. The proposed framework is readily extensible to accommodate any additional set of features. The next step is to add notes to the text about the areas of emotional interest. A "region of expressive interest" (RoEI) is a piece of speech that has a lot of emotional content. At the moment, labeling is based on the subjective empirical judgments of human speakers. However, it can also be done automatically using unsupervised grouping. The RoEIs play a certain role in the story because the speaker gave them that role or they help the story reach a short-term goal. They are also parts of

the speaker's storytelling style or method. The created online speech recognition and annotation tool mentioned in Section III is used to add notes to the RoEIs, generally at the word-level level. If these areas of emotional interest can be automatically found based on their traits using unstructured grouping, that would also be a very interesting study question. In order to study different types of expressive storytelling styles, this data-driven method aims to find patterns, regularities, and hidden structures in the speech. These can then be used to create a natural way of grouping expressive styles based on traits that can be measured. In contrast to the forced categorizations and a priori ideas that are common in expressive speech writings, this kind of emerging categorization would not only fit the data but would be driven by it. If you don't know ahead of time how many expressive categories (or clusters) are there, unsupervised Hierarchical Clustering (HC) is a good choice for the problem. For example, hierarchical k-means special cases of support vector machine Models and with Bayesian information criterion work well.

The (binary) hierarchy tree is a useful result of this method because it shows how clusters are formed in the data, how far apart they are, and how they gradually merge into bigger groups. With the help of the HC and the sounds of speech, an emerging emotional "typology" is created. It is important to keep in mind that this type of person is not based on or connected to any feeling theory. It just happened to appear in the data. For now, it would be enough for expressive speech synthesis to just "blindly" connect expressive categories to surface traits. Also, testing these groups on real people to see how well they make sense is very interesting.

You can think of the "non-expressive" part of the speech corpus as a different group. This is the speaker's "normal" or "neutral" way of talking. For modeling expressive categories, you can use the distribution of using the data's feature values to create statistical models of each expressive type. As with many other indications, speech's deviations and differences are sometimes more instructive than the values themselves. You would think that the difference between an accented word and plain speech would make its pitch and intensity stand out, rather than the absolute values that words obtain. Statistical models, particularly those concerned with generalizability, may benefit greatly by distinguishing between neutral and emotionally charged speech based on feature value differences.

You may also convert neutral speech to "colored" speech for every expressive category by comparing the statistical qualities of expressive categories to the characteristics of the typical speaking style. It all boils down to establishing a link between the necessary expressive group characteristics and those of neutral discourse. Synthesis technology, such as RANDOM FOREST or US, would have little trouble figuring out how to change vocal expressions. The next sections provide preliminary findings on the feature extraction and grouping approaches and annotation tools to aid with the processes and objectives we discussed before.

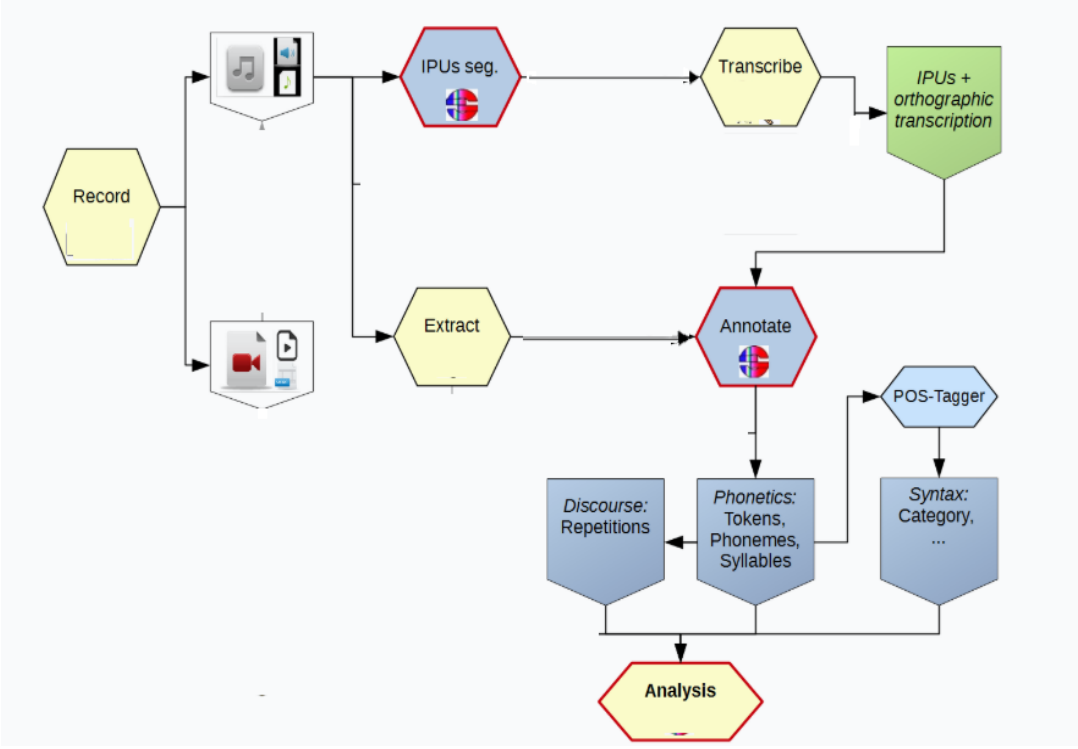

Fig 2. The experimental speech idea and marking tool for finding RoEIs includes: a) ways to define and set up a new experiment; and b) an example of labeling a sentence with flags that show how expressive each word is.

III. EXPERIMENTAL ENVIRONMENT FOR SPEECH PERCEPTION

The goal of the speech detection and labeling trial setting was to find examples of expressive / emotional behavior that make up RoEIs and then name them with a flag. Annotators can be both pros and amateurs when it comes to data marking. Listening to each sentence in the collection word by word and possibly giving each word a name is possible with this tool. When setting up an audio experiment, you can choose from basic flags and labels, or you can make your own creative categories. The process of adding notes to words is shown in figure 2.

III-Sound analysis, grouping, and outcomes

Building on the work in [4], an audiobook was now thought of as the spoken text to be analyzed. "A Room with a View," which was given by the Blizzard Challenge 2014 workshop data, was chosen as the podcast. As a result, it provides a good starting point for experiments because it includes about 8 hours of speech that was recorded in an emotionally aware way. The RoEIs are found and marked by hand in the voice data at the word level. People who heard the speech thought that certain words or groups of words carried important meaning or were used in a way that was meant to make people feel something. These words or groups of words are marked as RoEI. In this way, their function might be seen as closely connected to the dominance of the center. There are no clear marks or notes given; only simple on/off flags that show if a word is seen as expressively loaded.

Table 1 Acoustic Features Extracted at Word Level

Category	Feature Name	Description
Prosodic Features	Word Duration	Total time duration of the word (ms)
	Speaking Rate	Number of phonemes or syllables per second within the word
	Energy Mean	Average signal energy of the word
	Energy Variance	Variability of energy within the word

The traits used in this study have to do with speech numbers in time and space, such as pitch, strength, energy, and rate. The voiced frames of the word have been used to measure pitch and energy. The energy has been looked at by comparing the energy in the higher third of the frequency range to the total energy of the spectrum, with 100 being the starting point. Rate was taken into account by comparing the length of a phoneme to its average length across the whole speech collection. This is done by normalizing the length of each phoneme. It has been scaled so that a number of 100 means the rate is the same as the average rate of this phoneme in the speech collection. Table 1 is a summary of the set of sound features at the word level. The method used in this study is based on the idea that words are the basic units of expression, meaning that any emotional load put on speech changes whole words or groups of words. Pitch, passion, and energy have been found by analyzing each frame one at a time, but only for the voiced frames.

Rate has been worked out at the level of phonemes. To get numbers for these traits at the word level, the whole word's mean value (_MEAN) and standard deviation (_STD) have been found. The difference between a word's mean and variation values from the word before it (PREV_) and the word after it (NEXT_) has also been found.

Once all of the above audio quantities have been estimated, the next step is to use unsupervised hierarchical clustering methods (HC) in a series of tests to find the new expressive groups that the acoustic features capture. It is better to use hierarchical clustering instead of other common lustering methods like k-means clustering because you can't know ahead of time how many groups you want. The next step is to get statistical grouping (i.e., categories/class) descriptions for each possible category. In order to do some basic testing and try to answer the question of whether there is a "natural" way to divide the emotionally charged words into

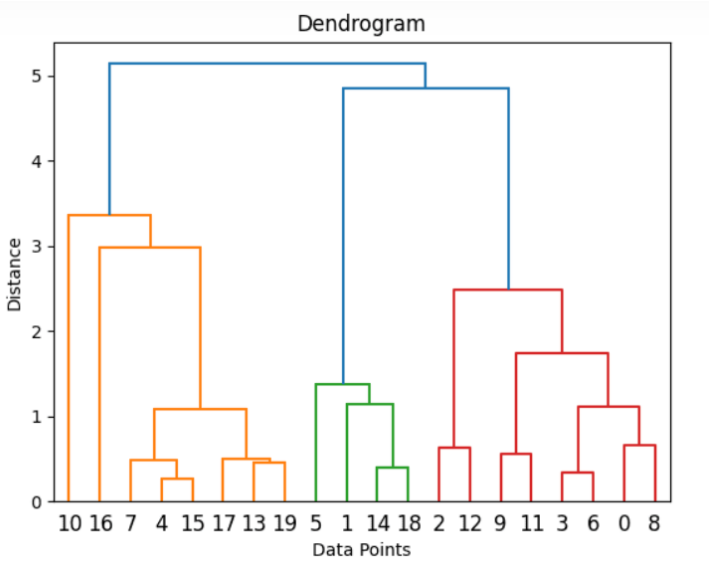

Fig 3. Some examples the word's given cluster or group is also shown as Ci, where i=2,4,...N and N is the number of clusters.

The outcomes of the HC analysis, categorized by their examined auditory properties, are shown in Figure 3. Ward's minimal variance approach and the HC analysis utilizes the Euclidean distance as the linkage criteria, which refers to the reduction in variance for the merged cluster. The HC yields a "clean" structure, characterized by well separated categories/clusters. By implementing a cut-off criterion on the distance between two merging clusters, a diverse array of categories arises. The picture illustrates an instance of prosodic differences between phrases of a speech, accompanied by designated labels as they arise from the HC processing. This functionality is integrated into the analytical experimental environment to provide enhanced visualization and oversight.

IV. CONCLUSIONS

We came up with a plan that includes speech analysis, modeling, and creation to make text-to-speech (TTS) systems that are very expressive. Unlike other methods, this system is based on core speech data and emotional speech patterns. This is different from the idea of fixed emotional models, which don't take into account the wide range of expressions and subtleties found in human speech. Stories and autobiographies are two popular types of writing that use this style.

Early findings from sound analysis and random grouping on a podcast dataset were shown to show how useful this method could be. In a story context, the goal was to find descriptive categories. In later work, different unsupervised hierarchical clustering algorithms, parameter change methods, ways to make expressive speech patterns that match the core data, emotional transitions, and temporal models will be looked into.

REFERENCES

1) Flanagan, J. L. (2013). Speech analysis synthesis and perception (Vol. 3). Springer Science & Business Media.
2) Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., &Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. Speech communication, 71, 10-49.
3) Fulop, S. A. (2011). Speech spectrum analysis. Springer Science & Business Media.
4) König, A., Satt, A., Sorin, A., Hoory, R., Toledo-Ronen, O., Derreumaux, A., ... & David, R. (2015). Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring, 1(1), 112-124.
5) Holube, I., Fredelake, S., Vlaming, M., &Kollmeier, B. (2010). Development and analysis of an international speech test signal (ISTS). International journal of audiology, 49(12), 891-903.
6) Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2011). Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. Journal of the royal society interface, 8(59), 842-855.
7) Shao, Y., Srinivasan, S., Jin, Z., & Wang, D. (2010). A computational auditory scene analysis system for speech segregation and robust speech recognition. Computer Speech & Language, 24(1), 77-93.

8) Boyce, S., Fell, H. J., & MacAuslan, J. (2012, September). SpeechMark: Landmark Detection Tool for Speech Analysis. In Interspeech (pp. 1894-1897).
9) Gangamohan, P., Kadiri, S. R., &Yegnanarayana, B. (2013, August). Analysis of emotional speech at subsegmental level. In Interspeech (Vol. 2013, pp. 1916-1920).
10) Akinwotu, S. A. (2013). A speech act analysis of the acceptance of nomination speeches of Chief Obafemi Awolowo and Chief MKO Abiola. English Linguistics Research, 2(1), 43-51.
11) Gaikwad, S. K., Gawali, B. W., &Yannawar, P. (2010). A review on speech recognition technique. International Journal of Computer Applications, 10(3), 16-24.
12) Simon, S., &Dejica-Cartis, D. (2015). Speech acts in written advertisements: Identification, classification and analysis. Procedia-Social and Behavioral Sciences, 192, 234-239.
13) Veaux, C., Yamagishi, J., & King, S. (2013, November). The voice bank corpus: Design, collection and data analysis of a large regional accent speech database. In 2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE) (pp. 1-4). IEEE.
14) Erjavec, K., & Kovačič, M. P. (2012). "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. Mass Communication and Society, 15(6), 899-920.
15) Chang, K. H., Fisher, D., Canny, J., & Hartmann, B. (2011, November). How's my mood and stress? An efficient speech analysis library for unobtrusive monitoring on mobile phones. In Proceedings of the 6th international conference on body area networks (pp. 71-77).
16) Van Eemeren, F. H., & Grootendorst, R. (2010). Speech acts in argumentative discussions: A theoretical model for the analysis of discussions directed towards solving conflicts of opinion (Vol. 1). Walter de Gruyter.
17) Stanig, P. (2015). Regulation of speech and media coverage of corruption: An empirical analysis of the Mexican press. American Journal of Political Science, 59(1), 175-193.