International Journal of Research
International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 4, April 2015
Available at http://internationaljournalofresearch.org

Automatic Language Identification System: A Review

[1]Niraj Kr.Singh & [2]Prof. Anoop Singh Poonia

[1]Research Scholar (M.Tech), [2]Professor, E & C Engineering
[1,2]Vivekananda Global University, Jaipur
[1]connect2vit@gmail.com, [2]pro.vcvgu@gmail.com

ABSTRACT:

Prosody is the part of speech where rhythm, stress, and intonation are reflected. In language identification tasks, these characteristics are assumed to be language dependent, and thus the language can be identified from them. In this paper, an automatic language recognition system that extracts prosody information from speech and makes decisions about the language with a generative classifier based on GMM is built.

INTRODUCTION

Evidences from prosody-based automatic language discrimination (LID) system suggest that the difficulties reported by other sites in incorporating prosodic information into LID systems may have been caused by their not using appropriate task-specific features. Running averages and correlations of prosodic features capturing syllable pitch and amplitude contours, duration and phrase location were evaluated by deriving a LLR function for each feature and language pair, then evaluating the effectiveness of that function as a discriminator. Data consists of speech in 11 languages (OGI database) representing a cross-section of traditional typological categories and relationships. Results show that prosody is highly useful in LID if complex perceptual events are broken down into simpler physical events and features are chosen based on task. Prosodic features can distinguish between language pairs as predicted by language typologies, suggesting that new languages can be classified using existing models of similar languages.

I. ISSUES IN LID SYSTEM

Other LID studies Past approaches to automatically identifying the language spoken in a conversational context have used broad phonetic features, detailed acoustic features, raw waveforms, pitch contours, vocabulary, etc. [4, 8]. The utility of prosodic cues like stress and rhythm realized as a function of three acoustic parameters (pitch, amplitude and duration) was unclear and therefore was typically not pursued in most studies. A few earlier attempts to use prosodic features found them only marginally successful: speech rate and syllable timing offered small improvements [5]; some differences were found between tone and non-tone languages in pitch change over the duration of the sentence and the word [6]. We would argue that prosodic features can be useful only if the appropriate features are used and that the lack of successful uses of such features in the earlier studies can be traced to not relying on task-appropriate features. It is not enough to derive a large set of general prosodic features because much more than language identity is encoded in the prosodic information. Suprasegmental features also encode discourse structure, emotion, native language and dialect, stylistics (e.g. read, spontaneous, lecture), utterance purpose (e.g. threaten, inform, persuade, flatter), speaker identity, etc. Since each aspect is encoded by a complex set of overlapping features, it is better to derive a smaller set that is maximally reliable for the task. A recent pairwise language discrimination study using only two prosodic features – F0 and amplitude envelope modulation – to discriminate between five languages with a recurrent neural

International Journal of Research
International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 4, April 2015
Available at http://internationaljournalofresearch.org

network has produced some of the most encouraging results for prosodic LID to date [1]. The network was able to find generalizations in the temporal patterns of the data; error distributions reflected traditional rhythm-related language classes. Our earlier work on pairwise discrimination between English, Spanish, Japanese and Mandarin used a much larger set of prosodic features than [1] and showed that those features can be very successful in LID [7]. We showed that the strengths of particular prosodic features and classes of features—primarily pitch, secondarily duration and location—reflect differences between the languages as predictable from prosodic classifications.

II. PROSODIC LANGUAGE CATEGORIES

The results from [1] and [7] suggest that a familiarity with the variation found in prosody and an understanding of the relationships between physical measurements and perceived events help in effectively identifying appropriate features, particularly if training data is limited, and predicting the discrimination success of specific language pairs. Most indepth cross-linguistic prosodic studies have focused on a small set of languages, on controlled speech, on particular theoretical claims, or are purely descriptive and the standard prosodic classification recognizes categories of pitch use (pitch-accent, tonal, non-tonal languages) and rhythm (syllable-timing, stresstiming, mora-timing). Pitch-related language categories differ with respect to amount of overall pitch variation, location of pitch change within a phrase, presence/absence of specific pitch contour types, pitch contours at different locations within a phrase, correlations between pitch and amplitude or duration features, and so on. Rhythm is crucial in parsing and intelligibility; however, there seems to be no simple measure of rhythm. Isochronous stresses or syllables are perceived and may be measurable in read speech or poetry reciting but are apparently not usually physically present in unplanned speech other than as tendencies. A solution is to break complex perceptual phenomena into simpler easy-to-measure interacting properties. A study comparing five languages differing in timing and tone concluded: “The difference between stress-timed and syllabletimed languages has to do with differences in syllable structure, vowel reduction, and the phonetic realization of stress and its influence on the linguistic system.” [2] The suggested existence of preferred tempos in the 1.4-2.0 Hz range [3] may also interact with syllable structure, vowel reduction and pitch use to explain a language's choice between salience of distance between syllables or stressed syllables. This suggests that by measuring simpler features such as distance between syllable onsets, between syllable nuclei, and between prominent stresses we should be able to identify differences between rhythmically different languages. In general, we expect languages that are more similar in pitch use and/or timing-related structure to be more difficult to differentiate automatically.

III. FEATURE VECTORS EXTRACTION

Many different algorithms exist for speech recognition and language identification. A common need between them is some form of parameterized representation of the speech input. These feature vector streams may then be used to train or interrogate the language models which will follow the feature extraction module in a typical language identification system.

It is obvious that there exist an infinite number of ways to encode the speech, depending upon which particular numerical measures are deemed useful. I will examine one feature extraction scheme which has been widely used. A block diagram of the scheme is shown below

This front-end has been studied by Davis and Mermelstein [15], and was shown to give the best performance of the options examined. Since then, it

has been used as a base for the comparison of different approaches to language identification by Zissman [16].

IV. CONCLUSION

In general, then, computationally efficient prosodic measures can provide a semi-independent noise resistant source of information for LID without any need for costly handtagging of training data. Our system could provide a very quick categorization of unknown language data. For best LID results, prosodic measures should be combined with other information, such as segmental distribution or word recognition. For discrimination in a multi-language context where the number and identity of the present languages are unknown, a good use of our fast prosody-based LID system can be to do an initial decision and a paring down of possible languages. Based on the prosodic categorization, an appropriately limited set of segment-based language models can be applied for the final LID decision.

REFERENCES

[1] D. Mart´ınez, O. Plchot, L. Burget, O. Glembek, P. Matejka, ˇ "Language identification in iVectors space", Proc. Interspeech 2011, Florence.

[2] N. Brummer, A. Strasheim, V. Hubeika, P. Mat ¨ ejka, P. ˇ Schwarz, J. Cernock ˇ y, "Discriminative acoustic language ´ recognition via channel-compensated GMM statistics", Proc. Interspeech 2009, Brighton.

[3] M.A. Zissman, "Comparison of four approaches to automatic language identification of telepone speech", IEEE Trans. Speech and Audio Proc., vol. 4, no. 1, pp. 31-44, 1996.

[4] T. Mikolov, O. Plchot, O. Glembek, P. Matejka, L. Burget, ˇ J. Cernock ˇ y, "PCA-based feature extraction for phonotactic ´ language recognition", Proc. Odyssey 2010 - The Speaker and Language Recognition Workshop, Brno.

[5] M. Soufifar, M. Kockmann, L. Burget, O. Plchot, O. Glembek, "iVector based approach to phonotactic language recognition", Proc. Interspeech 2011, Florence.

[6] L. Mary, B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition", Speech Communication 50 (2008) p. 782-796.

[7] Chi-Yueh Lin, Hsiao-Chuan Wang, "Language identification using pitch contour information", Proc. ICASSP 2005, Philadelphia.

[8] N. Dehak, P. Demouchel, P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification", IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 7, pp. 2095-2103, Sept. 2007.

[9] L.Ferrer, N. Scheffer, E. Shriberg, "A comparison of approaches for modeling prosodic features in speaker recognition", Proc. ICASSP 2010, Dallas.

[10] M. Kockmann, L. Burget, J. Cernock ˇ y, "Investigations into ´ prosodic syllable contour features for speaker recognition", Proc. ICASSP 2010, Dallas.

[11] M. Kockmann, L. Ferrer, L. Burget, and J. H. Cernock, "iVector fusion of prosodic and cepstral features for speaker verifi- cation", Proc. Interspeech 2011, Florence.

[12] M. Kockmann, L. Burget, O. Glembek, L. Ferrer, J. Cernock ˇ y, ´ "Prosodic speaker verification using subspace multinomial models with intersession compensation", Proc. Interspeech 2010, Makuhari.

[13] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-end factor analysis for speaker verification", IEEE Trans. on Audio, Speech and Language Processing, vol. 19, pp. 788-798, May 2011.

[14] S.B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", in IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-28, No. 4, August 1980.

[15] M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", in IEEE Transactions on Speech and Audio Processing, Vol. 4, No. 1, January 1996.