REVIEW OF AUTOMATIC LANGUAGE IDENTIFICATION

Pavel MATĚJKA, Doctoral Degree Programme (3)
Dept. of Radio Electronics, FEEC, BUT
E-mail: matejkap@feec.vutbr.cz

Supervised by: Doc. Ing. Milan Sigmund CSc., Doc. Dr. Ing. Jan Černocký

ABSTRACT

Language identification is an area that has had research interest form some time now, and there are some good papers that have been written on this subject. Just a few of these papers have been highlighted in this paper. They discuss some of the different ways in which language identification can be implemented.

1 REVIEW

Brief description of major studies and approaches to automatic language identification are described below.

1974-1980: Texas Instrument effort was based on frequency occurrence of certain reference sounds in different languages. Automatic segmentation to this reference sounds [1] brought results 64% correct on the test set consisting of seven languages. Further improvement were obtained in later studies by human interactive approach to determine reference sounds. The best published results [2] were 80% on five languages. Weakness of manuall determination of reference sounds was the main drawback of addition of another languages. This was shown in last paper [2] where seven languages task degrades performance from 72% to 62%. These studies embody notion of phonetic-base distinctiveness of languages.

1977: House and **Neuberg**'s work was based on manually phonetic transcribed data. HMM were trained on broad phonetic labels derived from phonetic transcription. They didn't use acoustic features. Their paper [3] showed perfect discrimination of eight languages and demonstrated that excellent language identification can be achieved by exploiting phonotactic information.

1980: Li and Edwars [4] applied Markov techniques suggested by House and Neuberg to real speech data. They used broad phonetic classes to compute two statistical models: one based on segments and other based on syllables. They reach 80% correct identification with five languages.

1982: Cimarusti and Ives designed a polynomial classifier on 100-elements LPC derived feature vector (including autocorrelation coefficients, cepstral coefficients, filter coefficients, log area ratios and formant frequencies) [5]. This approach was not based on phonetic segments but just on acoustic features. Overall accuracy 84% on eight languages demonstrates that language identification can be based only on acoustic features.

1986: Foil [6] examined two types of language identification system. First approach extracted seven prosodic features (based on rhythm and intonation) from pitch and energy contour. Second used formant frequencies (in terms of values and locations) to represent the characteristic sounds patterns of language. A k-means clustering algorithm and vector quantization were used. Language identification performance on three languages was 64% correct with 11% rejection on the data collected from radio with SNR 5dB.

1989: Goodman [7] extended Foil's work by modifying and adding parameters to feature vector and improving the classification distance metric.

1991: Sugiyama [8] performed vector quantization classification on LPC derived features. He explored the difference between using one VQ codebook per language vs. one common VQ codebook, in this case languages were classified according to their occurrence probability histogram patterns. The best overall recognition accuracy, 80% on 64 seconds of uknown speech, was obtained.

1992: Nakagawa [9] compared four methods - VQ (vector quantization), discrete HMM, continuous density HMM and GMM (Gaussian mixture distribution model). Comparative results for all mentioned methods were conducted on four languages. Results using continuous HMM and GMM (81.1%) were better than vector quantization (77.4%) and discrete HMM (47.6%).

1993: Muthusamy [10] dissertation on segmental approach to LID discusses which acoustic, broad phonetic and prosodic information is needed to achieve automatic LID. First experiments were conducted on 4 language task with high quality speech. On the basis of these promising results he further investigated this approaches with 10 language corpus of telephone fluently spoken speech [11]. Experiments with features based on pairs and triples of broad phonetic categories, spectral features (PLP) and pitch-based features were carried out on two languages (English vs. Japanese) and ten language task. The extension of frequency occurrence, segment ratios and duration were also explored. With system containing all above features merged together he got 48.5% on short utterances (avg. 13.4 sec) and 65.6% on long utterance (avg. 50 sec) on ten language task. He come up with conclusion that information on phonetic level instead of broad phonetic might be required to distinguish between languages with greater accuracy.

A perfect literature overview and multi-language speech corpus development [12] are great part of this work. Perceptual experiments were also conducted, in which trained listeners identified excerpts of speech of one-, two-, four- and six-second durations as one of the ten languages. The average performance over all languages rose from 37.0% to 43.0% to 51.2% to 54.6% with growing duration of speech.

1995: Yan [13] in his dissertation provided a partial unification by studying the roles of acoustic, phonotactic and prosodic information. Two novel information sources (backward LM and context-dependent duration model) were introduced. The best correct rates of 91% (45 second segments) and 77% (10 second segments) on nine language task were published. For the best system he used a set of six phone language-dependent recognizers based on HMM followed by language modeling of phone sequence for each language.

1996: Schultz et al [14] used large vocabulary continuous speech recognition system (LVCSR). They compared language identification system based on phone level and word level both with and without language model (LM). In the first attempt, bigram LM was implemented, but trigram in the second stage gained better results. Word-based system with trigram modeling of words (84%) outperformed the phone-based system with trigram modeling of phones (82.6%) significantly on four language task. They claim: The more knowledge is incorporated in the word-based language identification system, the better performance.

1999: Berkling [15] examined various ways to derive confidence measures for LID system. Three types of confidences were proposed. (1) Scores are polled according to winner – the target set contains scores of the correctly identified utterances. (2) Scores are pooled according to the input – the target set contains all scores where the input and the language model correspond to the same language. (3) Third method does not separate target and background but pools all winning scores into a single set regardless of whether or not the input utterance was correctly or incorrectly classified. He used phone recognizer followed by language models (PRLM) to evaluate which confidence measure is better. Experiments are conducted on NIST 1996 evaluation data. He used special measurement for evaluation. The Method 1 tracked the performance of the system best. He also studied adding new features (phone duration, phoneme frequency of occurrence ...) for improving confidence measure.

1999: Hombert and **Maddieson** [16] described using 'rare' segments for LID system. Detailed description of broad phonetic classes and its representation and behavior in different language families is provided, because segments which are rare and easy to identify are extremely valuable in an LID system.

2001: Navrátil [17] deals with a particularly successful approach based on phonotactic-acoustic features and presents systems for language identification as well as for unknown-language rejection. An architecture with multi-path decoding, improved phonotactic models using binary-tree structures and acoustic pronunciation models serve for discussion on these two tasks.

2002: Jayram et al [18] proposed a parallel sub-word recognition (PSWR) language identification system which is alternative to conventional Parallel Phone Recognition (PPR) system. Sub-Word Recognizer (SWR) is based on automatic segmentation followed by segment clustering and HMM modeling. PSWR outperformed PPR on six language task with 10 sec of testing utterance only on training set (90.2%) about 4% and it was 1% worse on testing set 62.3%.

2003: Adami [19] propose to used the temporal trajectories of fundamental frequency and short-term energy to segment and label the speech signal into a small set of

discrete units that are used for speaker and language identification. He derived new features with his own segmentation. He obtained 35% equal error rate on NIST 2003 LID evaluation with 30s utterances on 12 languages.

Adding duration of these 5 symbols decreased EER to 30%. He proved complementary information with phone-based system (24%) and fused these two systems he obtained 21.7%.

2003: MIT group [20] evaluated three approaches phone recognition, Gaussian mixture modeling and support vector machine classification and fusion of all above. They outline the differences and progress from the NIST evaluation in 1996 to the NIST evaluation in 2003. Main differences in GMM approach are using gender-dependent GMM and feature mapping techniques to channel independent feature space, in phone-based LID system new phoneme sets were used and trigram distributions were added to the language models, with language dependent weights for the trigrams, bigrams and unigrams.

2 CONCLUSION

Many approaches to implement automatic language identification were presented on many international conferences. There is one problem with comparison these approaches, because no standard database like (TIMIT) was available in the past. It becomes better, because National Institute of Standards and Technology (NIST)started a Language Identification Evaluations for comparative results all over the word and give support to new efforts in this field.

ACKNOWLEDGMENTS

This research has been partially supported by Grant Agency of Czech Republic under project No. 102/02/0124. Pavel Matějka was supported by doctoral grant of Grant Agency of Czech Republic No. 102/03/H105 and Jan Černocky by post-doctoral grant of Grant Agency of Czech Republic No. GA102/02/D108.

REFERENCES

[1] Leonard, R.G., Doddington, G.R. Automatic Language Identification. *Technical Report RADC-TR-74-200*, Air Force Rome Air Development Center, August 1974.

[2] Leonard, R.G. Language Recognition Test and Evaluation.. *Technical Report RADC-TR-80-83*, Air Force Rome Air Development Center, March 1980.

[3] House, A.S., Neuberg, E.P. Toward Automatic Identification of the Languages of an Utterance: Priliminary Methodological Considerations. *Journal of the Acoustical Society of America*, 62(3), pp. 708-713, 1977.

[4] Li, K.P., Edwards, T.J. Statistical Models for Automatic Language Identification. *Proc. ICASSP'80*, pp 884-887, April 1980.

[5] Cimarusti, D., Ives, R.B. Development of an Automatic Identification System of Spoken Languages: Phase 1. *Proc. ICASSP'82*, pp. 1661-1664, May 1982.

[6] Foil, J.T. Language Identification Using Noisy Speech, *Proc. ICASSP'86*, pp. 861-864, April 1986.

[7] Goodman, F.J., Martin, A.F., Wohlford, R.E. Improved Automatic Language Identification in Noisy Speech. *Proc. ICASSP'89*, pp. 528-531, May 1989.

[8] Sugiyama, M. Automatic Language recognition using acoustic features. *Proc. ICASSP'91*, pp. 813-816, May 1991.

[9] Nakagawa, S., Ueda, Y., Seino, T. Speaker-independent, Text-independent Language Identification by HMM. *Proc. ICSLP'92*, pp. 1011-1014, October 1992.

[10] Muthusamy, Y.K. *A Segmental Approach to Automatic Language Identification*. PhD thesis, Oregon Graduate Institue of Science and Technology, October 1993.

[11] OGI Multi Language Telephone Speech. *www.cslu.ogi.edu/corpora/mlts/*, Januar 2004.

[12] Muthusamy, Y.K. The OGI Multi-language Telephone Speech Corpus. *Proc. ICASSP'92*, pp. 892-897, October 1992.

[13] Yan, Y. *Development of an Approach to Language Identification Based on Language-dependent Phone Recognition*. PhD thesis, Oregon Graduate Institue of Science and Technology, October 1995.

[14] Schultz, T., Rogina, I., Waibel, A. LVCSR-Based Language Identification. *Proc. ICASSP'96*, pp. 781-784, May 1996.

[15] Berkling, K., Reynolds, D., Zissman, M.A. Evaluation of Confidence Measures for Language Identification. *Proc. Eurospeech'99*, vol. 1, pp. 363-366, September 1999.

[16] Hombert, J.M., Maddieson, I. The Use of 'Rare' Segments for Language Identification. *Proc. Eurospeech'99*, vol. 1, pp. 379-382, September 1999.

[17] Navrátil, J. Spoken Language Recognition–A step Toward Multilinguality in Speech Processing, *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 678-685, September 2001.

[18] Jayram, A.K.V.Sai, Ramasubramanian, V., Sreenivas, T.V. Automatic Language Recognition Using Acoustic Sub-word Units. *Proc. ICSLP'02*, pp. 81-84, 2002.

[19] Adami, A.G., Hermansky, H. Segmentation of Speech for Speaker and Language Recognition. *Proc. Eurospeech'03*, pp. 841-844, September 2003.

[20] Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., Reynolds, D.A. Acoustic,Phonetic,and Discriminative Approaches to Automatic Language Identification. *Proc. Eurospeech'03*, pp. 1345-1348, September 2003.