RANDOM WALK THEORY APPLIED TO LANGUAGE IDENTIFICATION

Etienne Marcheret *Michael Savic*
Electrical, Computer, and Systems Engineering Dept.
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

ABSTRACT

In this paper we discuss the most recent evaluation of the RPI language identification system by the National Institute of Standards and Technologies (NIST). This system is based on an acousto-phonetic approach where the phonemes present in a language are identified by a hidden semi-Markov model (HSMM). The HSMM was also developed at RPI. Knowledge of these phonemes provides us with the necessary probabilistic framework for classifier design. The classifier used in this system is designed in such a way that language specific scores generated during an evaluation form a random walk. Random walk theory has extensive applications in ecology, metallurgy, chemistry and physics. Until recently random walk theory has been primarily used as a tool for the measurement of the territory covered by a diffusing particle. We now show that random walk theory can be used to effectively design a language identification system.

1. INTRODUCTION

Over the past three years ([16], [11], [10], [9]) there has been considerable effort put into development of a language identification system by both government agencies and telecommunication companies. Although the needs of these two entities is quite different the impetus for both government and industry to seriously pursue the development of an automatic language identification system came from the availability of a large corpus of multi-lingual speech data. Beginning in the late 1980's and continuing today the Oregon Graduate Institute (OGI) has been collecting multi-lingual speech data. In 1993 the National Institute of Standards and Technologies (NIST) designated the OGI database as the standard for evaluating language identification algorithms. The standardized evaluation by NIST of existing automatic language identification systems has led to the development of a number of successful techniques. The basis for each of the techniques found in existing automatic language identification systems can be loosely categorized as acousto-phonetic, prosodic, phonotactic, or vocabulary. Acousto-phonetic systems rely on the phonetic inventory of a language, prosodics deal with phonetic duration and intonation of a language, phonotactics refer to the rules that govern the combinations of the different phones in a language, vocabulary refers to the word inventory of a language. Existing language identification systems based on the aforementioned categories can be found in literature: acousto-phonetic (([5]), ([7])), prosodic (([2]), ([6])), phonotactics (([16]), ([15])), large vocabulary continuous speech recognition ([8]). Accurate language identification systems have also been developed by combining multiple techniques ([4]).

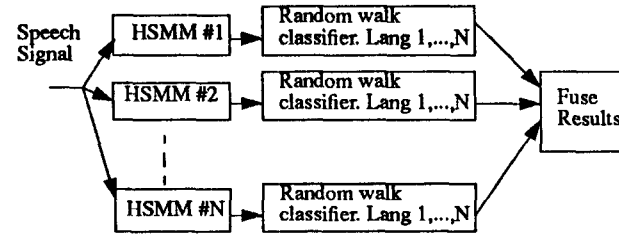

Figure 1: Language ID system, N languages.

In this paper we describe an automatic language identification system developed at RPI which is based on the acousto-phonetic technique. The novelty of this system is the technique used for classification. The classifier is designed in such a way that language specific scores generated during testing will form a random walk ([6]).

The remainder of this paper is formatted as follows: Section (2.) gives an overview of the system. Section (3.) describes random walk theory and how it fits into classifier design. Section (4.) present results generated by this system in comparison to existing systems from the NIST 1995 evaluation ([11]).

2. SYSTEM DESCRIPTION

Figure (1) shows the system architecture for N languages. The hidden semi-Markov model blocks are language specific phonetic segmenters. The design of each HSMM is accomplished using a 50 phoneme subset of the Worldbet phonetic transcription ([1]). The HSMM is an ergodic model and includes discrete durational probability modeling and continuous observation probability density modeling (mixed Gaussian). This HSMM was developed at RPI ([13]), and ([6]).

For each phoneme identified by each language specific HSMM a random walk classifier (section (3.)) was designed. The results of each random walk classifier are fused and a final decision is made on the language of the input speaker.

3. RANDOM WALK THEORY

Random walk theory provides a direct measure of the territory covered by a diffusing particle. Thus, this quantity appears in such fields as ecology, metallurgy, chemistry, and physics ([3]). This theory was used successfully in a speaker verification problem ([14]) and has now been used successfully for language identification. For each language specific HSMM and each phoneme a one dimensional random walk was formed for each language. This one dimensional random walk is generated by the normalized log likelihood ratio test:

$$\lambda'_{L_i}(\mathbf{x}_t) = \frac{\lambda_{L_i}(\mathbf{x}_t) - \mu_{L_i}}{\sigma_{L_i}}, \quad (1)$$

where the observation vector $\mathbf{x}_t$ is determined from a 16 dimensional cepstral vector calculated at time t. This cepstral vector is calculated at 12.5 msec intervals from a 25 msec frame of speech. The subscript L_i denotes language i. The log likelihood ratio test for each language is determined by a degenerate two class test:

$$\lambda_{L_i}(\mathbf{x}_t) = log \frac{Pr(L_i|\mathbf{x}_t)}{Pr(\bigcup_{n \neq i} n|\mathbf{x}_t)}. \quad (2)$$

The parameters μ_{L_i} and σ_{L_i} are determined such that if the incoming speech is language i the normalized log likelihood ratio test $\lambda'_{L_i}(\mathbf{x}_t)$ will produce a random variable with zero mean and unit variance (the parameters μ_{L_i} and σ_{L_i} are determined during training as a sample mean and sample variance respectively of the scores generated by equation (1) when the input speech is from language L_i). The language models $\lambda'_{L_n}(\mathbf{x}_t)$ where $n \neq i$ will produce a random variable having a non-zero mean, and a non-unit variance. Summation of the random variables $\lambda'_{L_i}(\mathbf{x}_t)$ over the number of observations (separately for each language model) we form the random walk result ([12]):

$$w_{L_i,j,t} = \sum_{\tau=0}^{t} \lambda'_{L_i,j}(\mathbf{x}_\tau). \quad (3)$$

where we have added the j subscript to denote the j^{th} phoneme. This accumulated result produced by the language model matching the language of the input speech will walk around the zero value (regardless of the number of λ'_{L_i}'s accumulated, this accumulated result will remain in a small neighborhood of the zero value). Applying the central limit theorem ([12]) we can conclude that the accumulated score result (equation (3)) for the model matching the language of the input speech will be governed by a zero mean Gaussian distribution where the variance will be a function of the number of observations used to form the summation (the variance will increase as the number of observations increase). This is given by equation (4).

$$f_{w_{L_i,j,t}}(x) = \frac{1}{\sqrt{2\pi t}} exp(-\frac{x^2}{2\sqrt{t}}). \quad (4)$$

The accumulated score results produced by the language models that do not match the language of the input speech will walk away from the zero value. Thus, the distance between the accumulated score result and the zero value will increase as the number of observations used to form the accumulated score result increases. Applying the central limit theorem the accumulated score results for the models that do not match the language of the input speech will be governed by a non-zero Gaussian distribution where the mean is an increasing function of the number of observations and the variance increases with the number of observations. This is given by equation (5).

$$f_{w_{L_k,j,t}}(x) = \frac{1}{\sigma\sqrt{2\pi t}} exp(-\frac{(x-\mu t)^2}{2\sigma\sqrt{t}}). \quad (5)$$

By combining the one dimensional random walks for each language model from each phoneme we form an n dimensional random walk $\mathbf{w}_{L_i,t}$ given by equation (6):

$$\mathbf{w}_{L_i,t} = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} t_1^{\frac{-1}{2}} \sum_{\tau=1}^{t_1} \lambda'_{i1}(\mathbf{x}(t_{1\tau})) \\ \vdots \\ t_n^{\frac{-1}{2}} \sum_{\tau=1}^{t_n} \lambda'_{in}(\mathbf{x}(t_{n\tau})) \end{bmatrix}. \quad (6)$$

Note that the accumulated score result from each phonetic segment j has been time normalized. The distribution of $\mathbf{w}_{L_i,t}$ will then be a multivariate Gaussian, which will remain at the origin for the correct language discriminant function, and will walk away from the origin for incorrect language discriminant functions. Thus, all that is needed to make a final decision about the language of the input speaker is a Euclidean metric which will tell us how far from the origin each language model has walked. The model that has produced scores that are closest to the origin is the model that best matches the language of the input speaker. This decision rule is given by equation (7):

$$\|\mathbf{w}_{L_i,t}\| < \|\mathbf{w}_{L_j,t}\| \ \forall \ j \neq i. \quad (7)$$

Figure (2) shows the resulting normalized score histogram for English and Mandarin models when the input language is English. The histogram has been determined from scores generated by 45 observations. Note that the histogram result follows closely to what would be predicted by the random walk theory (equations (4), and (5). Figure (3) shows normalized score histograms generated by the Mandarin model for the English speakers when 100 observations and 300 observations are available. We see that as the number of available observations increases the distance from the zero value increases (the mean of the distribution is an increasing function of the number of observations). The variance of the accumulated score distribution increases as well.

4. RESULTS OF THE 1995 NIST EVALUATION

Table (1) shows the results of the March 1995 NIST language identification evaluations. The results shown in this table were generated from English vs. Mandarin (MA), and English vs. Spanish (SP) for ten second speech segments and whole story speech segments. The whole story test required speech segments of more than 30 seconds.

NIST has also produced the ROC curves for an English vs. Spanish test including figure of merit measurements (FOM). These curves are shown in figure (4) for the ten

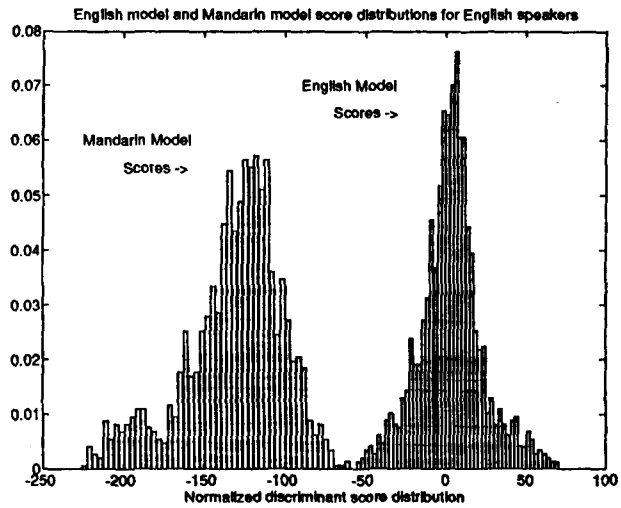

Figure 2: English model $\lambda'_{English}(\mathbf{x}_t)$ and Mandarin model $\lambda'_{Mandarin}(\mathbf{x}_t)$ score distributions for English speakers.

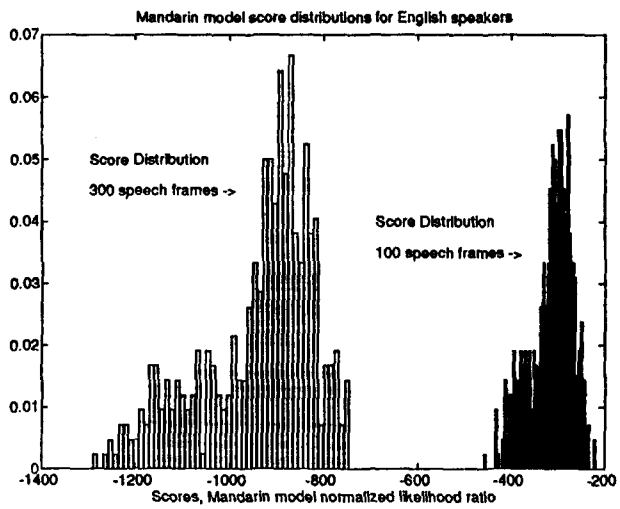

Figure 3: Mandarin model $\lambda'_{Mandarin}(\mathbf{x}_t)$ score distributions for English speakers, summations of 100 and 300 frames.

Table 1: 1995 Evaluation results from NIST, whole story segments (approximately 30 sec.), and 10 second segments (percent correct recognition).

Segment Duration	Whole Story		10 Second	
Site	MA	SP	MA	SP
AT&T	55	87	84	71
BBN	100	95	N/A	N/A
DRAGON	N/A	100	N/A	99
ITT	97	97	91	93
MIT - LL	97	87	97	86
LOCKHEED	63	100	66	95
NST	77	77	71	64
OGI	100	97	93	96
RPI	90	90	86	90

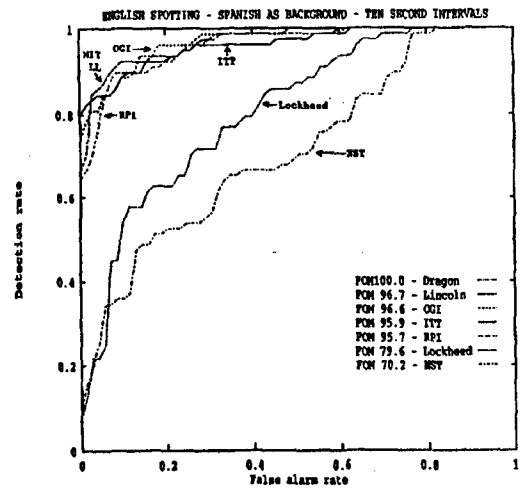

Figure 4: English vs. Spanish ROC curves, including FOM measurements for ten second segments.

second speech segment tests. Although the ROC curves tell us nothing about percent recognition accuracies they provide us with a measure of how separable the languages are to the system.

REFERENCES

[1] J. L. Hieronymus. Ascii phonetic symbols for the world's languages: Worldbet. *Journal of the International Phonetic Association*, 1993.

[2] S. E. Hutchins and Thyme-Gobbel. Experiments using prosody for language identification. In *Speech Research Symposium XIV*, pages 30–37, Johns Hopkins University, Baltimore MD, 1994.

[3] H. Larralde, P. Trunfio, S. Havlin, H. Stanley, and G. Weiss. Territory covered by n diffusing particles. *Nature*, 355, January 1992.

[4] K. P. Li. Experimental improvements of a language identification system. In *IEEE ICASSP*, volume 5, pages 3515–3518, May 1995.

[5] M. Lund and H. Gish. Language identification based on iterative estimation of language models. In *Speech Research Symposium XV*, pages 199–203, Johns Hopkins University, Baltimore MD, 1995.

[6] E. Marcheret. *New Approaches to Automatic Language Identification.* PhD thesis, Rensselaer Polytechnic Institute, May 1996.

[7] E. Marcheret and M. Savic. The rpi language identification system. In *Speech Research Symposium XV*, pages 84–88, Johns Hopkins University, Baltimore MD, 1995.

[8] S. Mendoza, M. Newman, Y. Ito, S. Lowe, and M. Mandel. Language idenitification through lvcsr. In *Speech Research Symposium XV*, pages 220–225, Johns Hopkins University, Baltimore MD, June 1995.

[9] N. I. of Standards and Technology. 1993 language identification protocols. In *First Language Identification Workshop*, Johns Hopkins University, Baltimore MD, June 1993.

[10] N. I. of Standards and Technology. 1994 language identification protocols. In *Second Language Identification Workshop*, Gaithersburg, Maryland, March 1994.

[11] N. I. of Standards and Technology. 1995 language identification protocols. In *Third Language Identification Workshop*, Gaithersburg, Maryland, March 1995.

[12] A. Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, 1984.

[13] K. B. N. Ratnayake. *New Approaches to Phoneme Based Speech Recognition.* PhD thesis, Rensselaer Polytechnic Institute, August 1993.

[14] J. Sorensen and M. Savic. Hierarchical pattern classification for high performance text-independent speaker verification systems. In *Speech Research Symposium XIII*, pages 49–52, Johns Hopkins University, Baltimore MD, June 1993.

[15] Y. Yan and E. Barnard. Recent improvements to a phonotactic approach to language identification. In *Speech Research Symposium XV*, pages 212–219, Johns Hopkins University, Baltimore MD, June 1995.

[16] M. Zissman and A. Martin. Language identification overview. In *Speech Research Symposium XV*, pages 2–14, Johns Hopkins University, Baltimore MD, June 1995.