STATISTICAL MODELS FOR AUTOMATIC LANGUAGE IDENTIFICATION

K. P. Li and T. J. Edwards

TRW
Defense and Space Systems Group
Redondo Beach, CA 90278

ABSTRACT

An Automatic Language Identification system simulation has been developed based upon an automatic acoustic-phonetic segmentation of speech. Utilizing six acoustic-phonetic segmentation classes, various finite-state models were developed to distinguish among five different languages. The finite-state models (trained with gathered segmentation language statistics) considered concatenations of individual segments as well as syllable-like strings. No attempt was made to locate syllable boundaries; therefore, the syllable models described either inter-syllable nuclei or intra-syllable nucleus segment statistics. Segmental durations were also included in some models. Language identification results ranged considerably across models, reaching a maximum of 80 percent correct identification for an independent test on 50 talkers (ten talkers per language).

INTRODUCTION

Several statistical models have been developed for an Automatic Language Identification (ALI) task. The basis for these models was an automatic segmentation which provided six acoustic-phonetic classes: (1) syllabic nuclei, (2) non-vowel sonorants, (3) vocal murmur, (4) voiced frication, (5) voiceless frication, and (6) silence and low energy segments. Based upon this segment data, two basic ALI statistical models were considered: segmental models and "syllable" models. The "syllable" models were further divided into two models, one of which utilized inter-syllable-nuclei segment sequences while the other utilized intra-syllable-nucleus segment sequences. The individual models were finite-state of zero and first order for the "syllable" models and zero through second order for the segment models. For some of the models, segmental duration information was included in the state diagram either as a dependent parameter or as an independent model.

Segmentation Data Base

The ALI front-end consisted of preprocessing followed by acoustic-phonetic segmentation as in Figure 1. Preprocessing included the extraction of several acoustic features such as voicing, zero-crossing count, and power. The segmentation then proceeded according to the following steps:

1. Classify speech/silence, voicing/voiceless, and frication/non-frication at a 10 msec frame rate.
2. Encode the frame into one of six aforementioned gross phonetic classes.
3. Concatenate adjacent frames with the same classification to form a segment.
4. Apply time-smoothing to eliminate extra or transitional segments with very short durations.
5. Find syllable nuclei/non-vowel sonorant boundaries to break up vocalic segments.

The segmentation data base for this study was the result of processing reading speech from twenty talkers from each of five different languages: two Asian and three Indo-European languages. All talkers were male, and dialect variations were allowed. The training data base consisted of 200 minutes (10 talkers x 4 minutes x 5 languages) of reading data resulting in roughly 42,000 syllables and 150,000 segments. The testing data base was half this size with 100 minutes (10x2x5) of reading speech.

Statistical Training

Each of the ALI finite-state models was either zero, first, or second order Markov in design. Using the training data, each model was obtained by gathering the proper finite-state statistics. The final result for each model was a state-table containing conditional (transition) probabilities.

Identification Strategy

ALI using the finite-state models proceeded by moving a variable "data" length analysis window through the training data and the independent test data. The analysis window was either x "segments" for the segment based models or y "syllables" for the syllable based models where "x" and "y" were varied to cover an analysis period of from fifteen seconds to two minutes. Each individual model was

tested over a selected analysis window with each language accumulating a conditional probability of being the language tested. Then, for each window, an accumulated weighted voting was obtained for each language based upon the conditional probabilities. The window was then incremented through the test data by one window element (segment or syllable) with new weighted votes accumulated until the data was exhausted for each talker. At this time, the language associated with the largest analysis-window vote for that talker was selected as the correct language.

SEGMENTAL MODELS

The segmental models were either zero, first, or second order Markov models. In addition, the distribution of segmental durations, the total amount of training and testing data required, talker variations, recognition strategies and scoring characteristics were studied. Table I presents an example comparing recognition results on training and independent test data for all three orders of the finite-state model. This confusion matrix shows the language identification result by the number of talkers. The following conclusions have been made based upon the segmental models' results:

1. Comparing recognition results between training and testing data, the higher order models show larger differences. This may indicate an insufficient training; i.e., additional talkers or speech data for training may be required. We found a drastic increase in recognition results when training on ten (70%) rather than five talkers per language (49%); however, prediction beyond ten talkers is not possible.
2. The higher order models provided better identification than the zero order, but the second order was not significantly improved over the first order results. Because the amount of training data is fixed, the second order model was not trained as well as the first order which may have resulted in a less significant improvement.
3. Typical of the results presented in Table 1, the Asian languages L4 and L5 are well separated from the Indo-European languages.
4. The longer the period of training and testing data provides a better and more stable result.
5. Talker variations or dialect variations within a language may be important to obtain a proper language model or multiple models.
6. When the segment duration is included in the model, a significant improvement occurred in the recognition of the training data (96%); however, the recognition on the testing data (62%) decreased perhaps due to a higher model dimensionality. Separate models incorporating only duration information performed very poorly (<40%).

SYLLABLE MODELS

The syllable models, as presented in Figure 2, were of two forms: inter-syllable and intra-syllable. The inter-syllable model was implemented only as zero-order Markov and described segment sequences between two syllabic nuclei (including the null sequence). The intra-syllable model described a "syllable" as a syllabic nuclei segment preceded by "p" and succeeded by "s" segments but not including neighboring syllabic nuclei. In Figure 2, "p" and "s" are both set to two segments. This model was implemented as both zero and first-order Markov.

The obvious reason for developing these "syllable" models was that these models do not require the detection of syllable boundaries (not available from our segmentation). In all, thirteen syllable models were evaluated including some incorporating "syllable durational" information as an independent model.

Tables 2 and 3 present typical results for the syllable models. In these figures the leftmost matrix presents the percentage recognition for each analysis window (10 syllables for Table 2 and 100 syllables for Table 3) while the rightmost matrix presents the number of talkers identified. The following conclusions were drawn from the syllable models data:

1. The syllable-like models converged to a useable number of unique segment sequence structures: 200-300 unique inter-syllable nuclei sequences and 370-450 unique intra-syllable sequences (p=s=2).
2. The required number of symmetrical intra-syllable segments (p,s) required to maximize recognition performance was two (p=s=2).
3. Truncating the syllable-sequence tables to the 100 most likely sequences per language improved recognition performance while decreasing memory. Purging the remaining table of non-language specific syllable structures reduced the table elements by one-half without affecting recognition performance.
4. The inter-syllable model (78%) is superior in performance for talker's language identification on independent test data.
5. The zero-order intra-syllable model (66%) did not perform as well as the inter-syllable model; but, the first-order Markov intra-syllable model with duration information, the data of which are presented in Table 3, provided the maximum performance attained on individual-window language identification for both testing and training data, but not for talker's language identification(56%).* However, the dimensionality of this latter model was 2401 as compared with 25 for the first-order segment model and 77 for the zero-order inter-syllable model. Additional training data are definitely required.
6. The inter-syllable model was better able to separate the Indo-European languages (L1-L3) from the Asian (L4,L5) than either the intra-syllable or the segment models.

---

* The higher percentages of individual-window ALI was due to talker's data being identified or not identified for the intra-syllable model whereas individual-window recognition for the inter-syllable model was a more "random" phenomenon.

7. A result similar to the segment models was obtained for talker variations and the requirement of a larger data base for higher order models.

Multi-Level Identification Strategy

A multi-level recognition strategy appears to be very promising with these models. The strategy calls for discriminating languages at the language family, group and individual language level as sub-problems. Preliminary results using an inter-syllable model at the language family level and segmental models at lower levels resulted in high 90% identification for the two family problem and an 80% average identification for the individual languages.

FURTHER RESEARCH

We are currently exploring ALI using the models developed here but on a much larger training data base. This training data base is nearly twice as large as currently used (approximately ten minutes per talker) and involves conversational rather than reading speech. The talker selection is also more realistic with nearly half of the subjects now female. Furthermore, the speech again involves telephone quality speech but now the speech is transmitted over an actual telephone channel rather than utilizing a 350-3300 Hz band pass filter as was used in the current study. This expanded study should provide an excellent test of the applicability of these models to useful, "real-world" ALI.

Table 1. Recognition results of zero, first, and second-order Markov Chain models for language identification using phonetic state sequences. Sub-decision window size is 100 segments.

	TOTAL	RECOGNITION ON TRAINING DATA			RECOGNITION ON TESTING DATA		
		Zero	First	Second	Zero	First	Second
L1	10	7	7	8	5	3	4
L2	10	1	7	7	6	10	10
L3	10	4	8	8	2	6	5
L4	10	7	10	10	6	8	8
L5	10	6	10	10	7	8	4
TOTAL	50	25	42	43	26	35	36

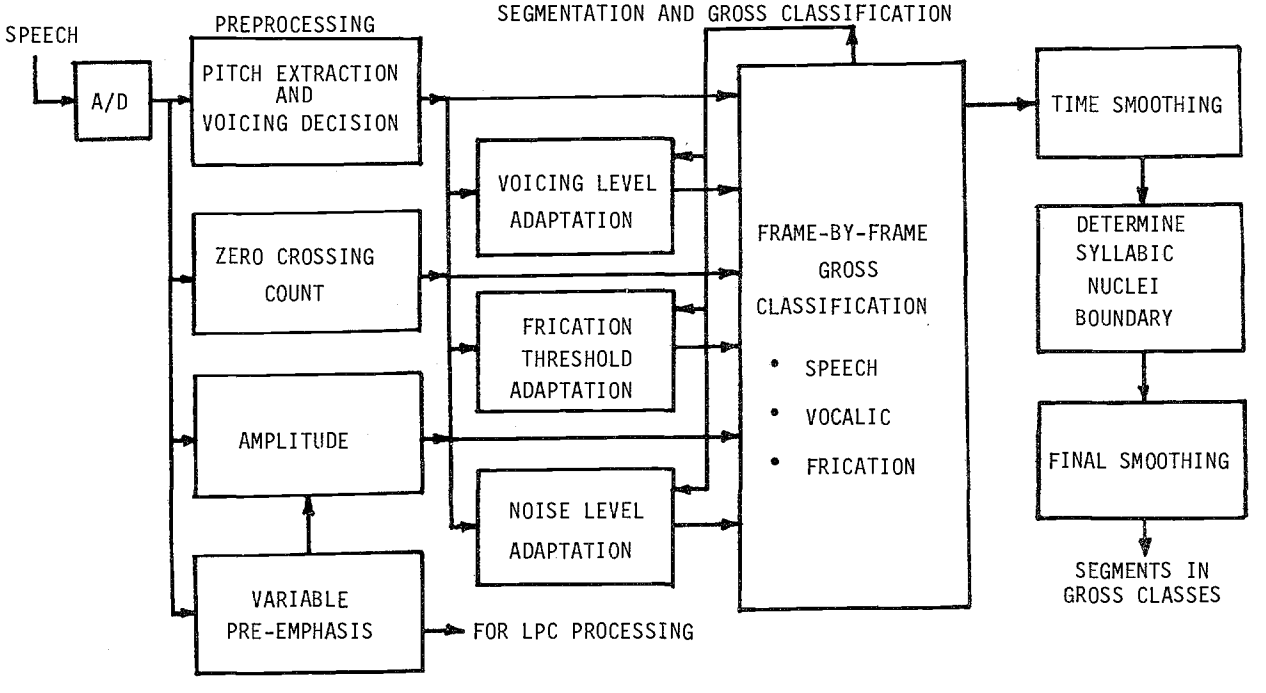

Figure 1. Block Diagram of Preprocessing, Segmentation and Gross Classification of Automatic Language Identification Experiments.

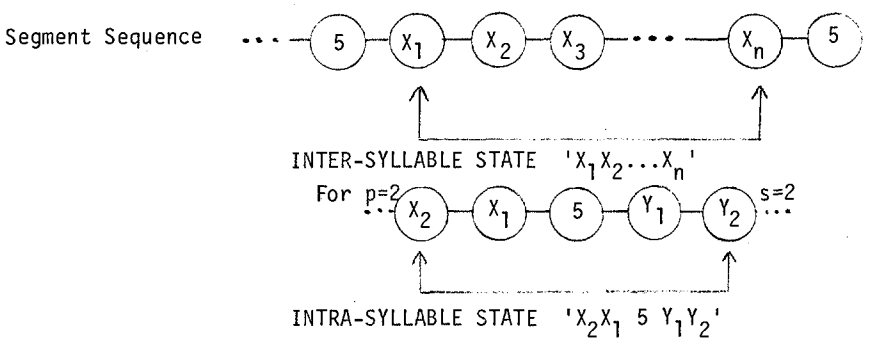

Figure 2. Two Syllable Parsing Strategies for Developing Language Models. Segment (5) Represents Syllable Nuclei. For Intra-Syllable State, All State Sequences Include a (5) State and are Terminated By Either a (5) State or the Maximum Number of States Allowed(p and s).

TABLE 2. Automatic language recognition results for the combined inter-syllable models. Confusion matrices for both training and testing data bases using a ten (10) syllable decision window. Each matrix row represents approximately 4000 sub-decisions.

Training Data

% Recognized as Language

Language Input	L1	L2	L3	L4	L5
L1	39	14	23	9	15
L2	21	31	23	11	13
L3	19	13	46	5	17
L4	7	7	6	62	18
L5	10	6	16	16	53

Recognized as Language

	L1	L2	L3	L4	L5
L1	7		3		
L2	2	6	1		1
L3			10		
L4				10	
L5			1		9

Independent Test Data

% Recognized as Language

Language Input	L1	L2	L3	L4	L5
L1	20	15	32	16	16
L2	10	39	24	16	11
L3	23	20	36	10	10
L4	7	12	15	48	18
L5	15	8	18	7	52

Recognized as Language

	L1	L2	L3	L4	L5
L1	2	1	7		
L2		9	1		
L3	1		9		
L4				9	1
L5					10

TABLE 3. Automatic language recognition results for the first-order Markov intra-syllable model with duration. Confusion matrices for both training and testing data bases using a 100 syllable decision window. Each matrix row represents approximately 4000 sub-decisions.

Training Data

% Recognized as Language

Language Input	L1	L2	L3	L4	L5
L1	92	4	1	0	2
L2	8	82	6	0	3
L3	2	2	94	0	2
L4	0	0	0	99	0
L5	4	1	0	0	95

Recognized as Language

	L1	L2	L3	L4	L5
L1	10				
L2		10			
L3			10		
L4				10	
L5					10

Independent Test Data

% Recognized as Language

Language Input	L1	L2	L3	L4	L5
L1	29	9	20	11	31
L2	19	42	20	3	16
L3	26	18	42	5	10
L4	3	11	14	53	19
L5	12	12	24	0	52

Recognized as Language

	L1	L2	L3	L4	L5
L1	4		1		5
L2	2	5	1		2
L3	3	1	5	1	
L4		1	2	7	
L5		1	2		7