# Identifying Language from Speech: An Example of High-Level, Statistically-Based Feature Extraction

Stan C. Kwasny
Center for Intelligent Computer Systems
Department of Computer Science
Washington University
St. Louis, Missouri 63130
(314) 935-6160
sck@cs.wustl.edu

Barry L. Kalman
Center for Intelligent Computer Systems
Department of Computer Science
Washington University
St. Louis, Missouri 63130
(314) 935-6160
barry@cs.wustl.edu

Weilan Wu
Center for Intelligent Computer Systems
Department of Computer Science
Washington University
St. Louis, Missouri 63130
(314) 935-6160
wwl@cs.wustl.edu

A. Maynard Engebretson
Central Institute for the Deaf
Department of Computer Science
Washington University
St. Louis, Missouri 63130
(314) 652-3200
ame@cs.wustl.edu

## Abstract

We are studying the extraction of high-level features of raw speech that are statistically-based. Given carefully chosen features, we conjecture that extraction can be performed reliably and in real time. As an example of this process, we demonstrate how speech samples can be classified reliably into categories according to what language was spoken.

The success of our method depends critically on the distributional patterns of speech over time. We observe that spoken communication among humans utilizes a myriad of devices to convey messages, including frequency, pitch, sequencing, etc., as well as prosodic and durational properties of the signal. The complexity of interactions among these are difficult to capture in any simplistic model which has necessitated the use of models capable of addressing this complexity, such as hidden Markov models and neural networks. We have chosen to use neural networks for this study.

A neural network is trained from speech samples collected from fluent, bilingual speakers in an anechoic chamber. These samples are classified according to what language is being spoken and randomly grouped into training and testing sets. Training is conducted over a fixed, short interval (segment) of speech, while testing involves applying the network multiple times to segments within a larger, variable-size window. Plurality vote determines the classification. Empirically, the proper size of the window can be chosen to yield virtually 100% classification accuracy for English and French in the tests we have performed.

## Introduction

In an international setting, one might overhear parts of conversations in a variety of languages. Given the proper experience, identifying familiar languages can be done easily and accurately. What is it that tells us the identity of a language? How do we know, for example, when the same speaker speaks English or French? Under the right circumstances, people seem to be able to tell immediately, often not from exactly what is being said, but from broad characteristics of the speech.

In fact, it is not necessary that one be competent in French to recognize that people are speaking French. When Arte Johnson speaks English with an accent, then suddenly starts talking in pseudo-German, the audience identifies the language as German, even though he may not use actual German words or phrases. It simply "sounds like German."

Spoken language is perceived on many levels. A variety of judgements about features of speech are constantly being made by a listener. Listeners unconsciously notice many things about speech -- tone of voice, style, pace, gender of the speaker, accent, degree of excitement, who is speaking, etc. These features can be very high level although often not consciously contemplated under ordinary circumstances by the listener. We further observe that spoken communication among humans utilizes a myriad of devices to convey messages, including frequency, pitch, and sequencing, as well as other prosodic and durational properties measurable in the signal. The complexity of interactions among these in the speech signal are impossible

to capture in any simplistic model necessitating the use of models such as hidden markov models and neural networks.

While speech understanding research has focused primarily on extracting "meaning" from speech, it is clear that there are many other ways humans process speech. Most of the high-level features mentioned above cannot be tied to any particular, conventional set of phonetic or acoustic features of the speech. Instead, they appear to be related to distributional patterns or statistical aggregates of the speech waveform.

We are investigating the extraction of high-level, statistically-based features from speech. Specifically, in this paper, the task is to determine the language being spoken from samples of raw speech. Bilingual speakers fluent in two languages are recorded and speech samples are separated into training and testing groups. Training attempts to create a network that can reliably determine which language is represented.

We assume that the classification task can be conducted in real time by the model. We further assume that it is only necessary for the model to see very raw speech waveforms, represented as sampled frequency bands over time. We specifically rule out explicit phonetic identification as well as a variety of other intermediate-level structuring that is typically found in speech understanding and recognition systems.

## Related Work

There have been several studies that demonstrate the existence of statistically significant differences among spoken languages at the acoustic level (Hanley, et al. (1966); Atkinson (1968)) and also at the level of phonetic features (Denes (1963); Kucera & Monroe (1968)). Abe et al. (1990; 1991) have considered some of the differences in automatically converting a speaker's voice from one language into another. Since these differences are measurable at the low end of the speech chain, then surely it must be possible to exploit those differences to build a model that emulates the human ability to correctly discriminate among languages.

House (1977) proposes a method of language identification which utilizes a language structure component in conjunction with a statistical component. His approach was apparently hindered, at that time, by the lack of sufficient computing power to perform the necessary statistical procedures.

We share some of House's beliefs about the value of statistical procedures in extracting certain high-level features. Being statistically based, the processing will naturally be resistant to noise and tolerant to some variation. We further assume that such processing can be demonstrated in real time. This assumption rules out the existence of a sophisticated language structure component and demands that intermediate levels of processing normally associated with speech understanding be finessed.

Recently, Muthusamy et al. (1990) has followed some of the suggestions made by House in examining this problem for four languages: American English, Japanese, Mandarin Chinese, and Tamil. They recorded six male and six female speakers each speaking 20 utterances in one of the languages. Four waveform and four spectral parameters were extracted and used to segment and label the speech with one of 7 broad phonetic categories with 82.3% accuracy. The segmented speech was then used in a second network designed to classify by language. This proved to be 79.3% accurate in classifying the speech into one of four languages.

Our approach differs from theirs in several respects. We first assume all processing can be conducted in real time. We also wish to finesse the need for intermediate structures as much as possible. We feel there is always some loss of information in mapping the waveform into discrete structures and this loss could have an effect on the success of the classification of the high-level feature.

## Data Collection

For our experiments, we collected speech samples from three bilingual speakers: two males and one female. All speakers fluently spoke English and one other language: $male_1$ spoke native French and non-native English; $male_2$ spoke native Japanese and non-native English; and $female_1$ spoke non-native French and native British English. Recordings were made of 12.5 second, randomly chosen samples of each speaker reading the phonetically balanced "rainbow passage" in English and excerpts of spoken passages read from newspaper stories in the other languages. Two different samples were recorded for each language for each speaker. Yielding a total of 12.5 speech samples.

All recordings were made in an anechoic chamber resulting in 16-bit samples at 24kHz. Five Band-Pass filters were used to separate the signal into bands which were low-pass filtered and decimated by a factor of 200. This process is illustrated in Figure 1.

Within the 12 second samples, we selected samples of smaller duration by specifying a start point and a duration and clipping it from the larger sample. This permits numerous overlapping samples to be extracted from each collected sample depending on the size of the sample to be extracted.

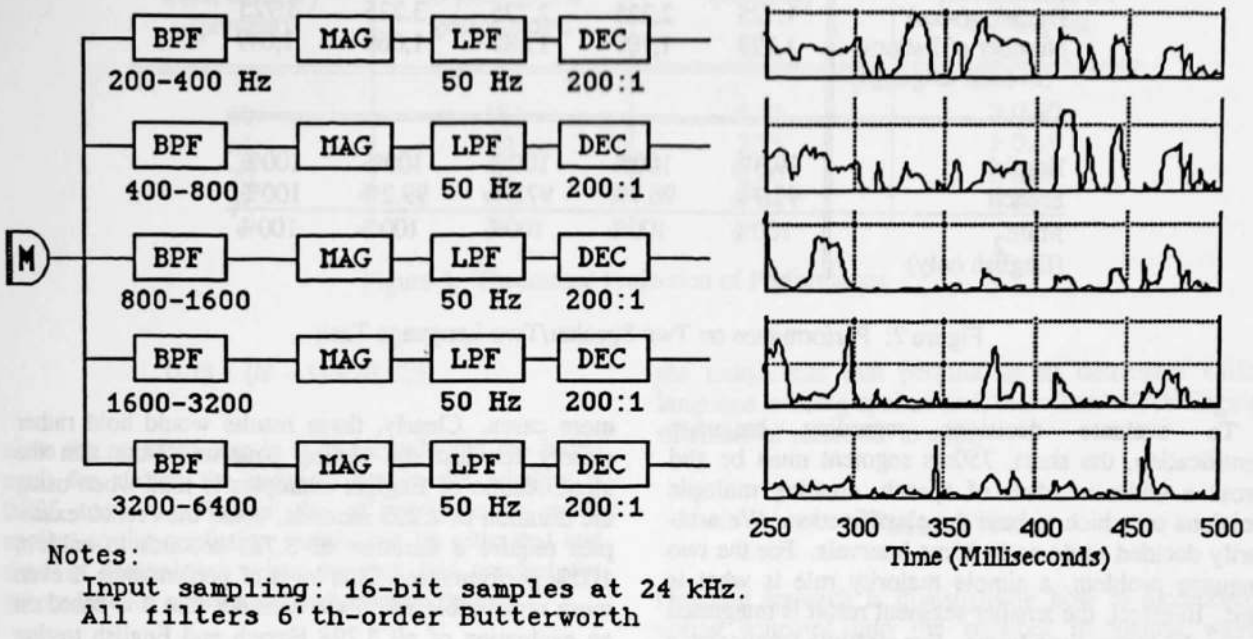


Figure 1: Frontend Processing

## System Design

Designing a system for this task requires that proper training is performed and that testing favors correct decisions. We are using a neural network that maps input units representing 750ms. duration of speech to output units representing the range of languages being identified.

The choice of 750ms is based on a compromise between a network that is too small to properly detect the distinctions necessary to identify the language and one that would require enormous computing resources to train. For durations less than 750ms, training patterns contain numerous input similarities which require separation as output dissimilarities. This can be determined by performing boundary pair testing as described in Kalman & Kwasny (1992). Such a situation is unacceptable since it indicates that good training will be extremely difficult to achieve. Durations above 750ms require an enormously large input layer and many network weights to manipulate. While a faster machine or more time could overcome such problems, we felt that this was also unacceptable for us given our current environment.

After choosing the size of the input window, the remainder of the architecture had to be determined. In our preliminary experiment, only two languages, French and English (spoken by $male_1$ and $female_1$), were used. Therefore, the output layer contains only two units, one for French and one for English. The size of the hidden layer is determined by making intelligent guesses. We examine the trainability of the network for all data, training and testing, and find the number of hidden units experimentally where the network maximally accounts for all the data. [Note that this number could also be found through a set of experiments in which training took place with just a randomly determined training set and then tested for generalization among the other patterns, but that method would take much longer.] The final network connects each layer to each layer forward of it, and so there are the standard layered connections as well as connections directly from the input to the output layer. All training was performed using variations on the conjugate gradient method (see Kalman, 1990 and Kalman & Kwasny, 1991).

During network training, its generalization capability is continually being monitored by calculating a confusion matrix for the testing set of patterns and applying a $\chi^2$ test to it. As the $\chi^2$ result continues to increase, training continues. If the test levels off or decreases, adjustments are made in training until the best trained network has been found.

| Repetitions | 40 | 60 | 80 | 100 | 120 |
|---|---|---|---|---|---|
| Threshold | 21 | 31 | 41 | 51 | 61 |
| Duration(secs) | 1.725 | 2.225 | 2.725 | 3.225 | 3.725 |
| Number of Patterns (in each language) | 1,729 | 1,709 | 1,689 | 1,669 | 1,649 |
| English | 99.3% | 100% | 100% | 100% | 100% |
| French | 92.7% | 96.7% | 97.3% | 99.2% | 100% |
| $Male_2$ (English only) | 100% | 100% | 100% | 100% | 100% |

Figure 2: Performance on Two Speaker/Two Language Task

To evaluate decisions regarding language identification, the short, 750ms segment must be slid across a wider window of speech, creating multiple decisions on which to base the classification. We arbitrarily decided to do so in 25ms intervals. For the two language problem, a simple majority rule is what is used. In effect, the smaller segment result is integrated across the larger time frame. For multiple languages, a plurality vote may be used and may potentially generate "don't know" classifications.

In analyzing the data by bands, the middle (third) band shares much with the adjacent bands. We decided to attempt to train the network from data further reduced by the elimination of band three. We successfully trained the network approximately the same level without including band three. This training is faster since there are fewer weights to adjust and so we used this method of training for all the results reported in the next section.

## Results

Our first results were obtained from experiments with two speakers, $male_1$ and $female_1$, each speaking English and French. While this is a very limited task, it represents the technique involved in successfully classifying speech segments for this purpose.

First, the 12.5 second speech samples of the two subjects were divided into training samples and testing samples. Each training sample was processed into 371 overlapping 750ms segments of speech each of which produced 360 numeric values of frequency information across the four bands (90 samples of 4 bands). Training proceeded to settle at 73.7% correct on the test patterns. This trained network was then evaluated on varying durations of windows and performance was measured according to a majority vote. Figure 2 shows the performance while varying the duration from 1.725 seconds to 3.725 seconds. We report figures on all data, both testing and training, to enable us to look at more cases. Clearly, these results would hold rather closely for just the testing patterns. Note that the identification of English examples is total when using the duration of 2.225 seconds, while the French examples require a duration of 3.725 seconds to achieve 100% performance. This level of performance is even more remarkable when we consider that it is based on an evaluation of all 3,298 French and English testing and training patterns.

We then tested the same network with English speech samples from $male_2$. These data are shown in the final row of Figure 2, with perfect performance achievable in a duration of about 1.725 seconds. This illustrates the degree to which the network is capable of generalizing to the speech of subjects for which it has not been trained, in this case $male_2$ whose native language is Japanese.

It is possible to make a theoretical analysis of the tradeoff between achieved performance level on the short segment and the duration of the window necessary for high-level performance (99.5% correct) during testing. Figure 3 shows such a theoretical projection for selected performance levels of the network. For example, if the network performs at the level of 60% correct for the worst category being classified, then assuming independent classificatory decisions (which is not strictly correct, but suitable for this approximation) we use the binomial theorem to yield

$$0.995 \leq \sum_{k=M}^{N} p^k (1-p)^{N-k} \binom{N}{k}$$

Here, N is assumed to be odd to make the calculation simpler, and M is assumed to be $\left\lfloor \frac{N}{2} \right\rfloor + 2$. So, in Figure 3, the initial column determines the probability, p, used in the binomial theorem and N is determined and shown in the second column. The third column can be derived from the second by the formula:

| Performance in worst category (percent) | Minumum (odd) N to yield 99.5% performance | Duration of window (seconds) | Normalized $\chi^2$ performance on training set |
|---|---|---|---|
| 60 | 181 | 5.25 | $\geq 0.40$ |
| 65 | 81 | 2.75 | $\geq 0.49$ |
| 70 | 51 | 2.00 | $\geq 0.55$ |
| 75 | 41 | 1.75 | $\geq 0.64$ |

Figure 3: Theoretical Projection of Performance

$$0.75 + (N - 1) \times (0.025)$$

since the segment size is 0.75 seconds and the increment for sliding the segment within the window is 0.025 seconds. Further the $\chi^2$ performance when applied to the confusion matrix can be estimated and used in determining when training has reached the proper level to achieve the performance desired.

## Conclusions

We have shown how a properly defined neural network is capable of reliably extracting the identification of what language is being spoken from raw speech. In our preliminary study reported here, perfect results were obtained by summing over multiple decisions and using a majority vote to determine a better decision from several individual error-prone ones. In fact, it can be shown that the error decays exponentially as the decision-making window is extended.

In a broader sense, we have illustrated the potential of extracting high-level features from raw speech by a majority decision-making system. The idea of "collecting votes" while sequentially processing input from a source channel is a powerful idea that results in noise tolerant decisions leading to remarkable performance. The majority vote technique exhibited here is a general method for improving the performance of an errorful method to one that is virtually flawless. Successful application of this method requires a task that submits to simple, aggregate classifications of the type demonstrated here and a classification technique that achieves a reasonable level of performance.

Human communication must carry information from one speaker to another by exploiting the characteristics of the channel. The channel of voice communication constrains what is permissible in a natural language utterance and what is not. Each language has developed its own unique system of utilizing the channel of communication to carry messages. While there is considerable overlap from language to language, it is the uniqueness that permits us to determine which language is being spoken and, therefore, which linguistic frame of reference to apply.

## Future Work

Ongoing research is investigating how to incorporate voting schemes into the network in natural ways. There is evidence for both spatial and temporal summation in nerve cells found in the brain and we hope to find architectures that better simulate such activity.

A promising approach involves the use of recurrent networks. In preliminary studies, a simple recurrent network was trained to achieve recognition rates competitive with those of non-recurrent ones, but using a much smaller window. Recurrent networks develop a limited memory of past events and can exhibit classification capabilities that consider both immediate inputs and past events.

Experiments have also begun which utilize the data we have collected to train networks for both gender discrimination and speaker discrimination. Here again, the thrust of the work is on reliable identification of high-level features in real time directly from the speech signal. With a small number of speakers, speaker discrimination is proving to be an easy task. This situation is expected to change as data from more speakers is collected. Our voting method is not expected to work quite as well with gender discrimination due to the large degree of overlap in vocal frequency between male and female speakers.

## References


Abe, M., and Shikano, K. 1991. Statistical analysis of bilingual speaker's speech for cross-language voice conversion. *Journal of the Acoustic Society of America* 90: 76-82.

Abe, M.; Shikano, K.; and Kuwabara, H. 1990. Cross-language voice conversion. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 345-348.

Atkinson, K. 1968. Language identification from non-segmental cues. *Journal of the Acoustic Society of America* 44, 378(A).

Denes, P.B. 1963. On the statistics of spoken English. *Journal of the Acoustic Society of America* 35, 892-904.

Hanley, T.D.; Snidecor, J.C.; and Ringel, R.L. 1966. Some acoustic difference among languages. *Phonetica* 14, 97-107.

House, A.S., and Neuberg, E.P. 1977. Toward automatic identification of the language of an utterance. I. Preliminary methodological consideration. *Journal of the Acoustic Society of America* 62(3), 708-713.

Kalman, B.L. 1990. Super Linear Learning in Back Propagation Neural Nets. Technical Report WUCS-90-21, Department of Computer Science, Washington University, St. Louis.

Kalman, Barry L., and Stan C. Kwasny. 1991. A superior error function for training neural networks. In Proceedings of the International Joint Conference on Neural Networks, Vol. 2, Seattle, Washington, 49–52.

Kalman, B.L., and S.C. Kwasny 1992. A training strategy for feed-forward neural networks based on input similarities. Submitted for publication.

Kucera, H., and Monroe, G.K. 1968. A comparative quantitative phonology of Russian, Czech, and German. New York: American Elsevier.

Muthusamy, Y.K.; Cole, R.A.; and Gopalakrishnan, M. 1991. A segment-based approach to automatic language identification In Proceedings of the 1991 IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto, Canada.