RECONCILIATION OF HUMAN AND MACHINE SPEECH RECOGNTITION PERFORMANCE

Misha Pavel[1]*, Malcolm Slaney*[2] *and Hynek Hermansky*[3]

[1]Biomedical Engineering, Oregon Health & Science University, pavelm@ohsu.edu
[2]Yahoo! Research, Santa Clara, CA, malcolm@ieee.org
[3]Electrical Engineering, Johns Hopkins University, hermansky@ieee.org

ABSTRACT

This paper focuses on resolving a number of issues that appear when the performance of human speech recognition is compared to that of automatic speech recognition. In particular human experimental data suggest that the resulting error is a product of the individual streams. On the other hand, Bayesian combination requires a multiplication of the estimates of prior probabilities and likelihoods. We show that, in principle, there is no discrepancy. The product of errors is a *performance* measure and human and machine performance may be consistent with this empirically established regularity. The product of probabilities is step in an algorithm to achieve the performance that may or may not be consistent with the product of errors. The main problem is that most of prior discussions failed to distinguish the performance measures from the estimates of the parameters used in the algorithm.

Index Terms— Speech Recogntion, Pattern Recogntion

1. INTRODUCTION

Despite major advances in automatic speech recognition (ASR), its performance usually falls short of the human ability to perceive and recognize speech. This discrepancy motivates investigations of human perceptual abilities as well as the differences between human and machine speech recognition with the ultimate goal to develop systems that exceed human performance. This approach requires a characterization of human performance, a complete description of the computational algorithms and an evaluation of the machine performance that can then be compared to those human listeners. It is useful to note that the performance of both human and ASR is greatly influenced by the contextual information. This information must, therefore be incorporated in both approaches.

Since in most realistic situations, the absolute performance of humans exceeds that of ASR, the Thus, it is possible to evaluate the relative improvement due to adding contextual information or the relative deterioration due to removal of portion of signals in different spectral bands. This approach enables one to estimate how efficacy of various information sources used in ASR. This approach is particularly attractive because of the availability of human performance data gathered during the last 100 years [**?**].

Our work is motivated by a discrepancy between the performance of a system based on machine-learning principles, and empirical models describing the performance of human listeners. This issue is important because recognizers combine information from multiple source. Simple probability models suggest the error rate should be related to the one minus the product of *correct* probabilities for each source, or perhaps given as the *sum* of error sources. Instead, it appears that human performance is a related to the *product* of errors. This paper aims to describe machine-learning models that can better model the (superior) human performance.

In this paper we summarize the results of experiments involving manipulations of the frequency content of the acoustic speech signals as well as the effect of additional contextual information on human performance. We then describe a general algorithmic approach used in ASR systems while clearly distinguishing the computational process from the performance measure of such systems. Finally we will illustrate situations where the performance of the algorithmic approaches is consistent with the human despite the different approaches to computation.

2. HUMAN PERFORMANCE CHARACTERIZATION

Human speech-recognition performance, intensely studied during the last 100 years, depends on a combination of acoustic information and on a variety of constraints derived from the context that limits the possible utterances. In this research, human subjects are confronted with acoustic stimuli and are asked to report what they heard – i.e. to recognize the speech sounds. Their performance is typically summarized by the proportion of correct responses; these summaries of the empirical results are the estimates of the theoretical representation of their performance denoted by the probability of correct responses, Q

Acoustic information is subject to a great deal of variability due to a variety of acoustic effects , ranging from specific

speakers' characteristics, e.g., accent, to the distortions due to background noise and room acoustics. One way that the human auditory system seems to cope with these effects involves dividing the acoustic input into frequency bands that are used as independent inputs to the recognition mechanisms; these are called critical bands. The discovery and recognition of this fact stimulated extensive experimentation by Fletcher and his colleagues, recently revived by Allen, provided an elegant summary of the interactions of information distributed over the acoustic frequency bands. The majority of these experiments were performed with meaningless, but pronounceable utterances, e.g., consonant-vowel-consonant sequences, chosen in order to minimize the ability of the listeners to use context, and in particular linguistic cues.

Although the mechanism underlying the combination process is not yet known in detail, Fletcher and his colleagues [**?**] discovered that the biological combination scheme appears to obey the following regularity. If the recognition performance to acoustic input X from several non-overlapping bands centered around frequency f_i with minimal context is given by $Q_{X,i}(X)$, then the human performance of the (acoustic-only) combination is approximately given by

$$Q_X(X) = 1 - \prod_{\forall i} (1 - Q_{X,i}(X)). \quad (1)$$

The implication of this functional form is that an error is committed only if none of the channels yields the correct classification.

As it turns out, a similar functional relationship holds when human listeners are aided by context, i.e., the human auditory system combines acoustic and contextual information sources. Contextual constraints can and frequently are expressed in terms of probability of various utterances. Context-related constraints can, for example, limit the number of possible words to two, e.g., yes or no,. Alternatively, context can restrict the possible words to those that pertain to a particular topic. Understanding the way that the human auditory system combines the acoustic and contextual information is likely to provide important insights into cognitive processes. This question stimulated behavioral research of the effect of context-related constraints on human recognition performance.

In particular, there is evidence [**?**, **?**, **?**] that the combination of contextual and acoustic effects can be characterzed in terms of the product of errors associated with these two separate types of information. In particular, if the performance of the human observer without acoustic input and based on guessing using only the context information is Q_C, then the performance due to the combination of context and acoustic inputs is approximately given by

$$Q(X) = 1 - (1 - Q_X(X))(1 - Q_C)^r \quad (2)$$

where $r > 0$ is a real scaling parameter, approximately usually between 1 and 2. In other words, the probability of an error is given by the product of the error due to acoustics times the probability of an error due to the contextual decision. Note that the probability of correct classification based on context is independent of the specific input signal. In this case, the effect of context represents the context-dependent prior probability of the correct class.

The product of errors as well as the fusion of context with acoustic information observed in human perception are both consistent with the notion of maximizing estimates of posterior class-probability as computed by Bayesian approaches. They represent ways to approximate Bayesian computations when the true probabilities are not available and the classifier has only access to the estimates of these probabilities. In the reminder of the paper we show that some of these behaviors are already consisted with the behavior of ASR systems. In particular in Section 3 we show the reason why an analogue of Equation 2 is a more efficient representation of fusion of acoustic and contextual information than a straight product. In Section 4 we demonstrate that a probabilistic representation of context and acoustic information can yield performance consistent with the product of errors. In Section 5 we note without proof that combing a large number of sub-optimal classifiers, consistent with the product of errors, can approximate the optimal performance obtainable by Bayesian representation. In conclusion we suggest that the approximations used by the human system may be a useful way to implement future ASR systems.

3. AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition (ASR) involves many subtle and complex computational algorithms that are well beyond the scope of this summary. For the purpose of this presentation, we focus on general aspects of the ASR that are common to most of the ASR systems.

Automatic speech recognition generally involves training and possibly validation sets of labeled acoustic examples. The training sets are assumed to be samples of acoustic utterances representing the actual situations and contexts that will confront the trained ASR systems. In addition to the acoustic samples, the training frequently involves textual training sets used to improve the estimates of prior probabilities of utterances. In either case, the proportion of examples in these training sets are assumed to represent the prior probabilities $\Pr\{L|C\}$ of labels L in context C. We note that the actual specification of the labels may involve complex models, sequences of models, e.g. utterances, associated with each label. We should also note that successful ASR systems may be trained on a variety of training sets. Moreover, in some situations, the prior probabilities may be determined from different training sets than those used to train the acoustics-based estimates. In any case, the acoustic measurements are assumed to be good estimates of the likelihood $\Pr\{X|L\}$. Unfortunately, ASR systems cannot compute these probabilities—they can

only output values that can at best be construed to be the estimates of these probabilities. We will denote the estimates obtained by the ASR system by $\hat{P}\{L|C\}$and $\hat{P}\{X|L\}$. We re-iterate that these quantities are only estimates of the actual probabilities and are themselves random variables whose distributions depend on the specific algorithms used to compute them and on the distribution of the input data.

The estimates of the prior probability and the likelihood are then used by the ASR system to estimate the posterior probability,

$$\hat{P}\{L|X,C\} = \frac{\hat{P}\{L,X|C\}}{\hat{P}\{X|C\}} = \frac{\hat{P}\{L|C\}\,\hat{P}\{X|L\}}{\sum_{i=1}^{M}\hat{P}\{X|L_i\}\,\hat{P}\{L_i|C\}} \tag{3}$$

where we assume a finite set of M labels L_i. The output determined by an ASR is usually chosen to be the most probable label, i.e., the label L^* that maximizes the estimate of the posterior probability:

$$L^* = \arg\max_{L}\left\{\hat{P}\left(L|X,C\right)\right\} \tag{4}$$

We note that in practice, most ASR systems estimate the logarithms of a quantity proportional to the probability of the observed data (without normalization), e.g., for the joint probability in Eq.(3),

$$\log\left(\hat{P}\{L,X|C\}\right) \approx \log\left[\hat{P}\{L|C\}\right] + \log\left[\hat{P}\{X|L\}\right] \tag{5}$$

where all three terms are random variables. The general assumption that holds in case of normal distributions is that the logarithms of the likelihood and prior probability estimates are random variables that represent unbiased estimates of the log probabilities. But Equation (5) represents essentially a linear regression problem, the left hand estimate should optimally be computed by weighted combination of the independent variables, i.e.,

$$\log\left(\hat{P}\{L,X|C\}\right) = a_1\log\left[\hat{P}\{X|L\}\right] + a_2\log\left[\hat{P}\{L|C\}\right], \tag{6}$$

where a_1, a_2 are real regression coefficients that depend inversely on the variance of the individual estimates. In particular, those inputs with higher variance will be associated with smaller weights. Since for the purpose of ASR, the absolute magnitude of the posterior probability is not essential, the quantity to be maximized in Equation (4) is actually

$$\lambda = \frac{\log\left(\hat{P}\{L,X|C\}\right)}{a_1} = \log\left[\hat{P}\{X|L\}\right] + \gamma\log\left[\hat{P}\{L|C\}\right] \tag{7}$$

where $\gamma = a_2/a_1$. Equivalently, the decision variable can be interpreted as an estimate of the posterior probability and is frequently written in terms of a product probabilities where the prior is raised to the power γ,

$$D = \hat{P}\{X|L\}\left(\hat{P}\{L|C\}\right)^{\gamma}. \tag{8}$$

We note that the decision variable in an ASR, i.e., D is based on a product of the estimates of the probabilities. This product, however, does not mean that the *performance* of the ASR cannot be consistent with the product of errors as shown in Equations (1) and (2). We will illustrate this phenomenon in the next section.

4. EFFECT OF CONTEXT — SIMPLE EXAMPLE

In this section we demonstrate that even a very simple ASR system that multiplies probability estimates to compute the value of the decision variable as in Eq. 5, the performance of the system can be consistent with the product of errors. Our system combines a simple model of acoustic recognition, with a simple class-based contextual model. The contextual model is always correct, up to a point. It knows the right class and the only errors it makes are because it doesn't know the specific instance in the class.

Assuming the approach to ASR given in Section 3, we examine a simple example of the effect of context on ASR performance. For the purpose of this example we assume that an ASR system is recognizing one of M labels as illustrated in Fig 4. Without loss of generality we assume that the correct response is L_1, and thus the probability of choosing L_1 given just the acoustic information is higher than all other choices, $\hat{P}\{L_1|X\} \geq \hat{P}\{L_i|X\}\,,\ \forall i > 1$. We further assume that if there is an error due to acoustics the erroneous labels are generated with probability $1/(M-1)$, a simplifying assumption.

The probability that the system assigns the correct label based only on the acoustics is equal to Q_X, but there is a context model which also must be considered. For this we assume that the correct context only allows labels $L \in [1,K]$, where $K < M$. This is equivalent to saying the context model always predicts the correct class (i.e. it's a vowel, or a verb) but it doesn't know which of the K labels is correct. Assuming all the labels consistent with the context are equally likely, the probability of a correct response based only on context is $Q_C = 1/K$. In effect this means that the contextual model produces a binary decision. It restricts the possible labels, but will not "fix" an acoustic mis-recognition.

When the context is combined with the acoustic analysis this simplified system assigns an incorrect label only if the acoustic analysis yields an error. The system has selected an incorrect response, but one that is consistent with the context, in this case labels $L \in [2,K]$. There are two cases that need to be considered: (1) the erroneous acoustic response is within the context set and (2) the erroneous acoustic response is outside the context set and the system must randomly select

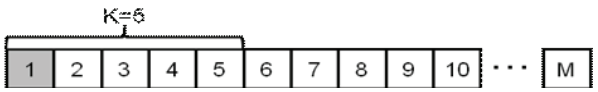

Fig. 1. A simple example of a integration of bottom up, acoustic recognition with a top-down, context based recognition. The numbers corresponds to the labels—the correct label is L=1. The context specifies K labels that are consistent with the context.

one of the remaining $K-1$ incorrect labels with probability $(K-1)/K$. Thus the probability of error with a system based on acoustics, X, and context, C, is the product of the acoustic error, times the two sources of error above, or

$$\begin{aligned} E(X,C) &= (1-Q_X)\left[g_C + (1-g_C)\left(\tfrac{K-1}{K}\right)\right] \\ &= (1-Q_X)\left(\tfrac{K-1}{K}\right)\left[\tfrac{M}{M-1}\right] \\ &= (1-Q_X)(1-Q_C)\left[\tfrac{M}{M-1}\right], \end{aligned} \tag{9}$$

where $g_C = \frac{K-1}{M-1}$ is the probability that acoustic response is within the set allowed by the (correct) context. The resulting system error is for all practical purposes approximates the product of acoustic and contextual errors, consitent with with the human empirical data, Eq. (2). A more complex analysis is required to determine the probability of error for more general distributions due to the context and acoustic estimates of the probabilities.

5. PERFORMANCE OF MULTI-STREAM ASR

In this final section we examine whether it is possible to build a detector that is consistent with the "product of errors" described by Eq. (1). It is easy to show that an optimal fusion algorithm combining conditionally-independent detectors that never make false-positive errors—as opposed to "no-response"—produces the desired result. This system makes an error only if none of the "high-threshold" detectors gives a response and the system has to make a guess. In this case the probability of making a error is given by the product of errors (no responses) of the individual detectors, in particular:

$$Q_X(X) = 1 - \frac{M-1}{M}\prod_{\forall i}\left(1 - Q_{X,i}(X)\right), \tag{10}$$

where M is the number of output classes. As M increases Eq. (10) approximates Eq. (1), which characterizes human performance.

In actual ASR systems, where the input is an acoustic signal X represented by real-valued features, the conditional distributions of the features, given class, are usually overlapping. In this case, detectors that do not make false alarm require "high thresholds" and yield suboptimal performance. It is, however, possible to approximate such a "high-threshold" system by combining a large number of conditionally independent, weaker classifiers or detectors [**?**]. When the number of these classifiers is large, the "high-threshold" detectors can be approximated with very simple combination rules such as a majority vote. Nevertheless, the resulting performance is not optimal in the sense of minimizing errors. This result, based on the notion of conditional independence of individual channels, suggests that a system that obeys Eq. (1) is suboptimal.

6. DISCUSSION

We demonstrated that although ASR algorithms compute a decision variables based on the product of the probability estimates, the performance of the system can be consistent with the product of error. We used a simple, but realistic model of acoustics and context (or language), yet when we analyzed the probability of incorrect recognition, the result in Eq. 9 is nearly identical to the model in Eq. 2 of how humans make errors. This is counterintuitive because one is maximizing performance by multiplying the estimates of probabilities, yet the system performance is described by multiplying one minus the probabilities. We demonstrated a simple implementation of a combination of acoustic with contextual constraints consistent with the biological data.

Interestingly, an implementation of a process that fuses independent channels that obey the product of errors rule turns out to be more difficult and controversial. It appears that the only way to specify such process involves suboptimal detectors and their combination. Although the individual high-threshold classifiers are suboptimal, as the number of classifiers increases, the combined error converges to zero and thereby approaches optimality. If this conjecture turns out to be true, we should be able to build machines that exceed the performance of humans.

7. REFERENCES

[1] Jont B. Allen, *Harvey Fletcher*, The ASA Edition of Speech and Hearing in Communication, published for the Acoustical Society of America by the American Institute of Physics, 1995.

[2] A. Boothroyd and S. Nittrouer, "Mathematical treatment of context effects in phoneme and word recognition," *J. Acoust. Soc. Am.*, vol. 84, no. 1, pp. 101–114, 1988.

[3] G. A. Miller, G. A. Heise, and W. Lichten, "The intelligibility of speech as a function of the context of the test material," *J. Exp. Psychol*, vol. 41, pp. 329–335, 1951.

[4] G. A. Miller, "Decision units in the perception of speech," *IRE Transactions on Information Theory*, vol. 81, no. 2, pp. 81–83, 1962.

[5] X. Song and M. Pavel, "Performance advantage of combined classifiers in multi-category cases: An analysis.," in *Neural Information Processing, 11th International Conference, ICONIP 2004*, Nikhil R. Pal, Nikola Kasabov, Rajani K. Mudi, Srimanta Pal, and Swapan K. Parui, Eds., Calcutta, India, 2004, Lecture Notes in Computer Science, pp. 750–757, Springer.