ON THE LIMITS OF SPEECH RECOGNITION IN NOISE

S. Douglas Peters[1], *Peter Stubley*[1,2] *& Jean-Marc Valin*[3]

1) Nortel Technology, 16 Place du Commerce, Montréal, Québec, CANADA H3E 1H6
2) l'Institut national de la recherche scientifique, 16 Place du Commerce, Montréal, Québec, CANADA H3E 1H6
3) Université de Sherbrooke, 2500, boulevard de l'Université Sherbrooke, Québec, CANADA J1K 2R1

ABSTRACT

In this article, we consider the performance of speech recognition in noise and focus on its sensitivity to the acoustic feature set. In particular, we examine the perceived information reduction imposed on a speech signal using a feature extraction method commonly used for automatic speech recognition. We observe that the human recognition rates on noisy digit strings drop considerably as the speech signal undergoes the typical loss of phase and loss of frequency resolution. Steps are taken to ensure that human subjects are constrained in ways similar to that of an automatic recognizer. The high correlation between the performance of the human listeners and that of our connected digit recognizer leads us to some interesting conclusions, including that typical cepstral processing is insufficient to support speech information in noise.

1. INTRODUCTION

Automatic speech recognition (ASR) has made considerable strides in the last number of decades. This is particularly evident in the fact that ASR products are currently performing such tasks as dictation, voice-dialing, directory assistance and automated attendant functions. However, these advances are due in large part to the advances made in computing technology. ASR is still hampered by the challenges that exist in the areas of robustness to speaker, channel, and especially background noise. In 1997, Shigeki Sagayama released the results of a informal poll of ASR technology managers and researchers identifying robustness to noise as the most significant limitation of existing technology [1]. It is in this context that we address the problem of speech recognition in noise.

In the last decade, a large number of methods have been proposed in the literature to deal with the problem of speech recognition in noisy environments. For example, there is a family of methods based on HMM decomposition including PMC and NOVO [2], [3], [4]. These methods have distinct models for both speech and noise which are jointly matched to a given token. Another family of methods comes from Carnegie Mellon, and includes RATZ, STAR, VTS, and a number of varations on the cepstral normalization theme (see, for example, [5] and its references). All of these methods manipulate the features or the models (sharing the same feature space) under the assumption that the conventional cepstral representation of speech is sufficient for the purposes of robust speech recognition. In this article, we challenge this assumption in an oblique way by asking the question, "How would the information reduction inherent in standard mel-cepstral processing affect the ability of human beings to recognize speech?"

In order to answer this question, we have constructed an experiment that "levels the field" between human and machine. That is, we asked a number of human listeners to identify the content of speech that had been processed in a manner consistent with mel-cepstral processing. The details of this processing will be discussed in a subsequent section. Before we proceed, however, it is of interest to describe some of the other steps taken in order to isolate the findings of this study to the information capture of an acoustic front-end.

In most respects, humans quite naturally have the advantage over the machine: our syntactic and semantic processing mechanisms are in considerable advance of the most sophisticated automated systems. In order to mitigate this disparity, we have chosen for our study a very simple lexicon and grammar that can be modeled precisely in the on-line search: connected strings of digits. This grammar can be visualized as in Figure 1 below:

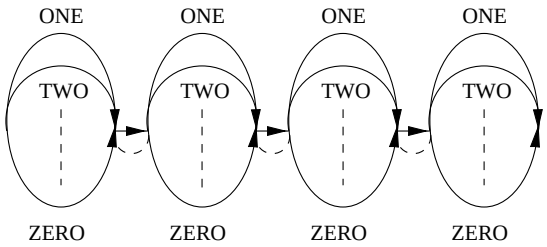

Figure 1. Connected digit grammar

On the other hand, the computer has two significant advantages over humans in the context of this experiment. First, the acoustic models are constructed explicitly in the target space. That is, there is a training mismatch for the human, who models speech based on full-bandwidth, full-resolution data. Second, the computer has perfect memory, and will not "forget" the early digits in a string by the time it is listening to the final digits in that string. These disparities were also addressed in the present experiment. For every processing considered, the human subjects were permitted a training phase in which they could warm to the unfamiliar speech representation. Further, the tokens considered were all four digits long, reducing the cognitive load, and subjects were permitted to replay any given token as many times as they pleased before making a recognition decision.

This study is different from previous investigations into the limits of human speech recognition in two respects. First, this study makes use of an important and meaningful simple grammar. Second, and more significantly, we expose our listening subjects to audio stimuli that has been corrupted in keeping with mel-frequency cepstral processing. That is, the humans essentially hear the same thing that machines "hear".

After a brief description of the corpus of noisy speech under consideration, the varieties of processing that were used

in our experiment will be described in detail. In Section 4, further experimental details will be addressed. An automatic speech recognizer whose performance is compared to that of humans will then be described, and Sections 6 and 7 will present the experimental results and a detailed classification and regression tree analysis, respectively. The final section will consist of discussion and conclusions.

2. DATABASE

A database of tokens was provided by one hundred speakers who made calls from the passenger seat of various cars touring around Montreal and its associated highways. The analog cellular phones in use were all operating in hands-free mode, a microphone having been attached to the visor above the front passenger seat of the car in which the speaker remained for the duration of his contribution. Each speaker provided approximately fifty tokens from a predefined list of randomly generated four digit strings.

The contributions of eighty of the speakers were used to adapt existing word-specific gender-dependent connected-digit models. The remaining twenty speakers (almost 1000 tokens) provided the test set for this experiment. These data were sampled at 8kHz and μ-law quantized in the manner typical for toll-quality audio.

3. PROCESSING

3.1. Forward processing

A feature extraction process typical for ASR systems is illustrated in Figure 2.

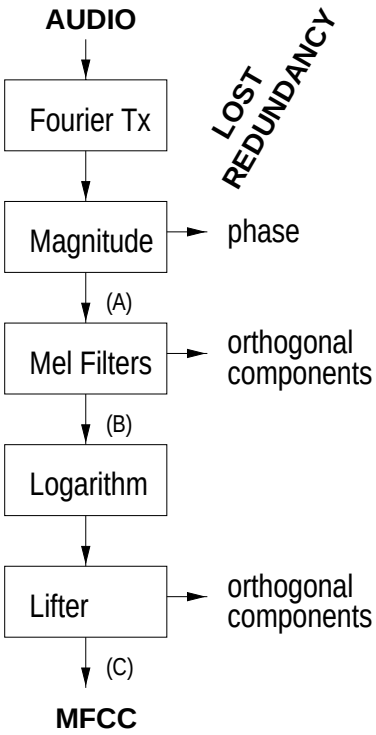

Figure 2. Feature extraction

In general, the spectral resolution reduction and phase elimination characteristic of this feature extraction is common to nearly every ASR front-end in existence. Historically, the justification for these processing steps is as follows. First, there is evidence to suggest that humans are relatively insensitive to phase [6]. Second, typical Fourier analysis is simply too complex for modeling purposes, and resolution reduction follows. Critical-band filtering is used by analogy to human inner ear excitation models. Alternatively, linear prediction is used by analogy to the vocal tract speech production mechanism. In either case, considerable loss in resolution results. Finally, a further reduction in spectral resolution is due to the cepstral liftering, whose original motivation involved the elimination of pitch information.

The processing implicit in Figure 2 can also be expressed mathematically as

$$\mathbf{c}_k = \mathbf{D}\ log\left[max\left(\mathbf{H}|\mathbf{F}\mathbf{x}_k|^2, \epsilon\right)\right]. \qquad (1)$$

Here, ϵ is a vector of floors for the frame channel energies. These channel energies are simply the product of the real-valued channel filter matrix $\mathbf{H}$ and the power spectrum for the k^{th} frame, $|\mathbf{F}\mathbf{x}_k|^2$. The complex-valued matrix $\mathbf{F}$ represents the Discrete Fourier Transform. The matrix $\mathbf{D}$ is typically a subset of the Discrete Cosine Transform, resulting in real-valued cepstral features $\mathbf{c}$. In the present case, the matrix $\mathbf{D}$ has eight columns, corresponding to the low-index cepstral coefficients. The matrix $\mathbf{H}$ has twenty rows, corresponding to twenty triangular mel-frequency filters, and 128 columns, in keeping with 256-pt FFT's. Further, the Fourier analysis used Hanning windows on 25ms frames which were overlapped by 50% with the adjacent frames.

3.2. Inverse processing

Given the pre-processing definition captured in Figure 2, one can reconstruct an audio signal from any level of processing. We considered four different processing mechanisms, three of which are marked as (A), (B), and (C) to correspond to the depth of processing indicated in Figure 2. These can be defined mathematically as:

$$\begin{aligned}
\mathbf{x}_{(A),k} &= \mathbf{F}^{-1}\left\{\sqrt{|\mathbf{F}\mathbf{x}_k|^2}e^{j\phi_k}\right\} \\
\mathbf{x}_{(B),k} &= \mathbf{F}^{-1}\left\{\sqrt{max\left[\mathbf{H}^{\#}\mathbf{H}|\mathbf{F}\mathbf{x}_k|^2, \mathbf{0}\right]}e^{j\phi_k}\right\} \\
\mathbf{x}_{(C),k} &= \mathbf{F}^{-1}\left\{\sqrt{max\left[\mathbf{H}^{\#}exp\left(\mathbf{D}^{\#}\mathbf{c}_k\right), \mathbf{0}\right]}e^{j\phi_k}\right\}.
\end{aligned}$$

Here, ϕ_k is a random phase vector, and the superscript $\#$ denotes pseudo-inverse, i.e., $\mathbf{A}^{\#} \triangleq \mathbf{A}^T\left(\mathbf{A}\mathbf{A}^T\right)^{-1}$. This pseudo-inverse results in a minimum-norm transformation from the lower-dimensional space back into the space of higher dimensionality. A final processing was considered, which we will label as (D):

$$\mathbf{x}_{(D),k} = \mathbf{F}^{-1}\left\{\sqrt{max\left[\mathbf{H}^{\#}exp\left(\mathbf{D}^{\#}\mathbf{c}_k\right), \mathbf{0}\right]}e^{j\angle(\mathbf{F}\mathbf{x}_k)}\right\}.$$

where $\angle$ denotes the vector of angles of the complex vector argument. In effect, processing (D) feeds forward the original phase of the signal. While the construction of a signal with high-resolution phase and low-resolution amplitude is certainly questionable, it was felt that this processing would provide a feeling for the effects of amplitude resolution reduction without a complete loss of phase information. Finally, the audio sequence was reconstructed from the resulting frame audio vectors $\mathbf{x}_k$ using a method analogous to overlap-and-add convolution.

4. EXPERIMENTS

A tool was constructed to deliver processed-audio tokens to a workstation user. Fifteen human subjects participated in the experiment. These subjects were encouraged to use this tool as much as possible over a period of three weeks in

sessions giving as many as fifty tokens per session. At the start of each sesssion, the tool chose a random processing and also randomized the test list. The subject was then encouraged to become familiar with the type of processing determined for that session by listening to a number of test tokens. When the subject considered themselves sufficiently familiar with the sound of the processed speech, they would then begin the official part of the session, providing the tool with a recognized four-digit string on the keypad of their workstation. More than 2100 human recognitions were logged.

5. RECOGNIZER

Since the object of this research is ultimately to improve the performance of automatic speech recognizers, it is of interest to compare the human recognition to that of an ASR system. In principle, the ASR pre-processor has already been defined in Section 3. However, there is a significant difference between the ways in which a human and a computer handles the temporal aspects of speech. The human is clearly capable of modeling static and dynamic elements of speech in a consistent manner and of adapting those models to any token-dependent anomalies that may arise. On the other hand, the best-to-date ASR systems still use an inconsistent modeling of the dynamic aspects of speech by augmenting the feature vector with the so-called "delta" coefficients. Moreover, the best that a computer can do to deal with token-dependent biases is to apply some simple conditioning to its features such as cepstral mean subtraction (CMS).

The pre-processing used for our ASR system is therefore equivalent to that denoted (C). That is, eight cepstral coefficients are used. The length of the feature vectors used for this experiment, however, is sixteen. That is, eight "delta" parameters, approximating the time derivatives of the cepstral coefficients at an appropriate modulation frequency are also used. Further, the energy parameter (zero-th cepstral coefficient) is replaced by its "delta-delta", or approximate second derivative value. Finally, the effects of bias on the cepstral parameters are removed be a standard CMS technique. We will denote the processing of the ASR system (M) simply as a convenient way to distinguish the performance of that system from that of the humans on the informationally equivalent processing (C).

A large corpus of wireless connected-digit tokens were used to train word-specific, gender-dependent HMM models in the defined feature space. These models were then adapted to noisy hands-free data that were similar but independent from the test set, as previously mentioned in Section 2. These word models used between nine and twelve states and a generous number of Gaussians in mixture models for observation probability density functions. Silence models were also trained on clean and noisy wireless data, and were optionally skipped in the recognizer search.

6. RESULTS

The overall results are summarized in Table 1. The digit and string recognition rates (DRR and SRR, respectively) are shown for each of the processings considered. It is of immediate interest that the performance of the ASR system is so close to that of the human subjects under processing (C). While more detailed discussion will be given in a later section, the clear suggestion of this result is that the weakness of our ASR sytem in noise is related primarily to the pre-processing. The effect of any weaknesses in acoustic modeling or noise compensation seem to be very much less significant than the limitations of the features themselves.

Of course, there are many factors in play in this experiment. For example, there could be considerable variability between speakers, token singal-to-noise ratios (SNR's), and also listener's capabilities. In order to investigate this matter further, the recognition data were assessed using decision tree analysis.

Table 1. Recognition summary

Processing	tokens	DRR	SRR
(A)	475	97.1%	91.5%
(B)	569	92.3%	77.3%
(C)	661	89.3%	71.4%
(D)	451	92.1%	78.3%
(M)	998	86.9%	65.0%

7. ANALYSIS

A decision tree is a method to assess the most significant factor related to the variability of an observable quantity [7]. The observable quantity for present purposes is the digit recognition rate. The factors provided to the decision tree as candidates for data splitting included

- Speaker (S): {a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t}
- Speaker gender (G): {M,F}
- Digit (D): {1,2,3,4,5,6,7,8,9,0}
- Processing (P): {A,B,C,D}
- SNR (R): (-10,30)
- Listener (L): {a,b,c,d,e,f,g,h,i,j,k,l,m,n,o}
- Position (X): {1,2,3,4}

The decision tree then selects the factor for which a data split results in the greatest reduction of total distortion of the partitioned observable. The first such test was performed on the entire collection of human recognition data, and is pictured in Figure 3.

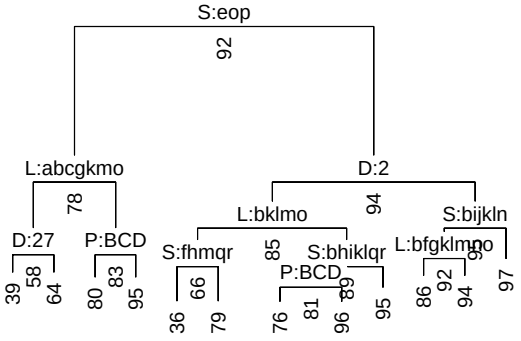

Figure 3. Global HSR decision tree

In this figure, each decision is illustrated by a horizontal line, and the significance of that decision is indicated by the length of the vertical lines on either side. The decision is encoded according to the factor of significance, which is shown above the center of the horizontal line. The digit recognition rate of each partition is indicated at each node of the tree. As is evident from this figure, the most significant factor related to human digit misrecognition is the speaker. In particular, three speakers result in a much lower recognition rate than the remaining seventeen. As it happens, the tokens from the three speakers in question were acquired under particularly noisy conditions. The road noise for these tokens was nonstationary, as in the case when a car is travelling over a blocked concrete highway or when

there is a resonance associated with the wind at a partially open window. Note, however, that the token-wide SNR for these tokens was not significantly higher than that for other tokens.

On the next level of decision, in the domain of the tokens with the difficult noise, the significant factor is found to be the listener. In fact, the grouping of the listeners, with a few exceptions, was on the basis of first language. Almost half of the experiment's listening subjects speak English as a second language. For difficult speakers and adapted (native) listeners, the processing is the next most significant factor. In fact, for these difficult speakers, any of the low-information processings results in a more than four-fold increase in the digit error rate. This indicates that a reduction in the spectral resolution renders the signal insufficient to carry speech information in the context of the noisy channel under consideration, even for native listeners.

A summary of the human recognition performance partitioned with respect to speaker (noise) and listener category is given in Table 2 below. The first number in each central entry is the human DRR under processing (A), and the second number is the DRR for all other processings.

Table 2. HSR breakdown

	stationary noise	nonstationary noise
native listener	97.4%/94.6%	95.1%/79.5%
nonnative listener	97.4%/90.5%	not enough tokens

It is interesting to note from this table that the digit recognition rate for the well-behaved conditions and processing (A) is the same for both categories of listener. However, under the more difficult processing regimes, the nonnative listeners make twice the number of errors as the native listeners. Similarly, either more difficult conditions or more difficult processing results in a doubling of the error rate for the native listeners. The combination of difficult conditions and difficult processing, however, makes the error rate increase by a factor of eight.

A second data frame was constructed in which only processing (C), which is informationally equivalent to that of the ASR system, was considered. Moreover, the ASR system itself was introduced as another "listener" with code M. The resulting decision tree is shown in Figure 4.

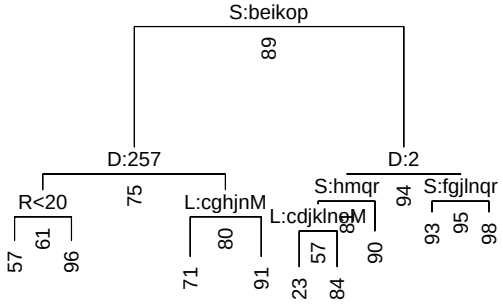

Figure 4. SR decision tree with (C/M)

Here, we observe that once again the speaker is the most significant factor partitioning the data on the basis of digit recognition rate. In this case, however, three other speakers were associated with those considered difficult.

Finally, it is interesting to note that in no cases was the recognizer isolated by a decision but was rather associated with the nonnative listeners.

8. DISCUSSION & CONCLUSION

In this paper, the results of an experiment examining human speech recognition in noise were presented. The experiment largely mitigated the effects of human language modeling. More interestingly, the human subjects were exposed to audio for which information was reduced in a manner equivalent to standard MFCC processing. As a result, the subjects were able to listen to, among other processings, the same MFCC's that were "heard" by a typical automatic recognizer.

In summary, it was found that there was a considerable increase in recognition error rate at each stage of processing. The elimination of phase consistent with the calculation of the power spectrum, for example, reduced the recognition rate from almost 100% to 91.5%. This observation suggests that any "phase-deafness" on the part of humans is only in effect in a steady-state context. The degradation in human recognition rate was even more significant, however, with the resolution reduction inherent in the calculation of mel-frequency filters and the subsequent cepstral liftering.

In decision tree analysis, it was discovered that the most significant factor in the loss of recognition was the noise environment. The native language of the listener was also a significant factor. This second effect suggests that there is a subtle modeling process in human recognition between what ASR researchers typically define as acoustic modeling and language modeling. Native listeners have captured this likely segmental effect in their model of speech whereas nonnative listeners are insufficiently adapted. It is also interesting that analysis consistently puts the ASR system among the nonnative listeners as insufficiently adapted to this phenomenon.

There is no question that there is a large amount of redundancy in speech. Moreover, in a high capacity channel with little noise it is likely that the traditional cepstrally liftered features, where a great deal of this redundant information is lost, are sufficient to carry the speech information. However, when the channel loses capacity due to noise, more redundancy is required. This claim is borne out by the current results which demonstrate at least a doubling of the human digit recognition error rate for cepstral-equivalent processing in stationary low-SNR conditions, and a four-fold increase in the nonstationary low-SNR case.

REFERENCES

[1] Sagayama S. and Kiyoami A. (1997) "Issues relating the future of ASR for telecommunications applications" *Proc. ETRW '97*, Pont-à-Mousson, France, pp. 75–81.

[2] Varga A.P and Moore R.K. (1990) "Hidden Markov model decomposition of speech and noise," *Proc. ICASSP '90*, Albuquerque, pp. 845–848.

[3] Gales M.J.F. and Young S.J. (1992) "An improved approach to the hidden Markov model decomposition of speech and noise," *Proc. ICASSP'92*, San Francisco, pp. 233–236.

[4] Martin F. et al., (1992) "Recognition of noisy speech by using the composition of hidden Markov models," *Proc. ASJ Conf. '92*, pp. 7–10.

[5] Stern R.M., Raj B. and Moreno P.J. (1997) "Compensation for environmental degradation in automatic speech recognition," *Proc. ETRW '97*, pp. 33–42.

[6] Deller J.R., Proakis J.G. and Hansen J.H.L. (1993), *Discrete-Time Processing of Speech Signals*, MacMillan.

[7] Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. (1984) *Classification and Regression Trees*, Wadsworth and Brooks.