A MULTI-EXPERT SPEECH RECOGNITION SYSTEM USING ACOUSTIC AND MYOELECTRIC SIGNALS

A. D. C. Chan[1,2], K. Englehart[1,2], B. Hudgins[1,2], D. F. Lovely[2]
[1]Institute of Biomedical Engineering, University of New Brunswick, NB, Canada
[2]Department of Electrical and Computer Engineering, University of New Brunswick, NB, Canada

INTRODUCTION

Automatic speech recognition (ASR) is a potential alternative control technology for high performance jet aircraft. ASR can improve pilot efficiency and safety by simplifying the user interface and encouraging "head-up" flying [1]. Unfortunately, conditions during flight in a jet aircraft are not ideal for conventional ASR systems, which only use acoustic speech information to perform speech recognition. High ambient noise within the cockpit and various stress conditions that a pilot must endure while flying (e.g high G-force, positive pressure breathing, vibration) degrade the classification accuracy of conventional ASR systems [1].

Recently, we proposed using the myoelectric signal (MES) from articulatory muscles of the face as a second source of speech information [2,3,4]. In this paper, a new method of combining the opinions of experts is presented. This method is based on evidence theory. Experimental results demonstrate that using this method of combining the opinions of experts a MES ASR expert can effectively complement an acoustic ASR expert, resulting a multi-expert ASR system that is resilient to noise

EVIDENCE THEORY

Evidence theory was developed by Dempster [5] and later refined by Shafer [6]. They introduced a mathematical framework that enables the precise assignment of partial beliefs to sets of classes. These partial beliefs can be used to compute two dual nonadditive measures: belief and plausibility. Dempster's rule of combination also provides a method of combining partial beliefs from distinct bodies of evidence, which can be used to combine the opinions from multiple experts.

Frame of discernment

Assume there are N classes, denoted C_i (i = 1, 2, …, N). In the context of ASR, these classes would correspond to words in the ASR system's vocabulary. Let the set of classes be denoted as $\Theta = \{C_i: i = 1, 2, \ldots, N\}$, which is known as the frame of discernment. Define 2^Θ to be the power set of Θ, which is the set of all subsets of Θ, including the whole set Θ and the empty set $\varnothing$. In evidence theory, we are concerned with propositions that the truth or correct class is in the set $A \in 2^\Theta$.

Basic probability assignment

A basic probability assignment (BPA) m(A) is a function that precisely assigns a portion of belief to a set $A \in 2^\Theta$. A BPA has the constraints that the range of the BPA is between zero and one, the BPA assigns no portion of belief to the empty set, and the sum of all portions of belief assigned by the BPA is equal to 1. Mathematically these constraints can be stated as:

1. $m(A) \in [0,1]$ (1)

2. $m(\varnothing) = 0$ (2)

3. $\sum_{A \subseteq \Theta} m(A) = 1$ (3)

Belief and plausibility

The belief measure for a set $A \in 2^\Theta$ can be computed from the BPA using the formula:

$$Bel(A) = \sum_{B|B \subseteq A} m(B) \tag{4}$$

The plausibility measure for a set $A \in 2^\Theta$ can be computed from the BPA using the formula:

$$Pl(A) = \sum_{B|B \cap A \neq \varnothing} m(B) \tag{5}$$

The belief measure is the sum of all portions of belief assigned to the set A and all subsets of A. The belief measure sums the portions of belief assuming all uncertainties do not support the proposition A, as it does not include portions of belief assigned to sets that partially intersect A. The plausibility measure is the sum of all portions of belief that overlap the set A. The plausibility measure sums the portions of belief assuming all uncertainties support the proposition A. It sums all the portions of belief used in the belief measure and also includes portions of belief that partially overlap the set A. In this paper we will be using the plausibility measure.

Dempster's rule of combination

Evidence from two independent bodies of evidence with BPAs m_1 and m_2 can be combined using Dempster's rule of combination:

$$m_{1,2}(A) = \frac{\sum_{B \cap C = A} m_1(B) m_2(C)}{1 - \sum_{B \cap C = \varnothing} m_1(B) m_2(C)} \quad \text{if } A \neq \varnothing \qquad (6a)$$

$$m_{1,2}(A) = \varnothing \quad \text{if } A = \varnothing \qquad (6b)$$

According to the numerator of equation 6a, if the first body of evidence assigns a portion of belief $m_1(B)$ to proposition B and the second body of evidence assigns a portion of belief $m_2(C)$ to proposition C, these beliefs are combined by taking the product $m_1(B)m_2(C)$ and assigns this portion of belief to the intersection of the propositions A = $B \cap C$. If proposition B and C do not overlap their intersection is empty $B \cap C = \varnothing$. The denominator of equation 6a is a scaling factor that the sum of all portions of belief assigned is equal to one (BPA constraint #3, equation 3) by accounting for portions of belief that would have been assigned to empty intersections.

If there are more than two bodies of evidence, they can be combined by applying Dempster's rule of combination iteratively. The order in which the bodies of evidence are combined does not affect the result.

COMBINING THE OPINIONS OF EXPERTS

Assume there are N classes, denoted C_i ($i \in \Lambda = \{1, 2, \ldots, N\}$) and the set of classes is denoted $\Theta = \{C_i: i \in \Lambda\}$, as in the previous section. Also, assume there are M experts denoted e_j (j = 1, 2, …, M). For a given classification sample $x \in C_n \in \Theta$, expert e_j produces a score vector $S_j = [s_j(1)\ s_j(2) \ldots s_j(N)]$, where $s_j(i)$ is a measure of the degree of the confidence that the expert believes $x \in C_i$. Define the rank vector $R_j = [r_j(1)\ r_j(2) \ldots r_j(N)]$, such that $r_j(i) \in \Lambda$ is the index of the class with the i^{th} highest score assigned by expert e_j. Expert e_j can independently classify x by choosing the class with the highest score ($e_j(x) = C_k$, $k = r_j(1)$)

In a multi-expert system E, a method of combining the opinions or scores from multiple experts is required to choose a single class.

Borda count

A traditional method of combining the opinion of experts is using Borda count [7]. For a given classification sample $x \in C_k$, expert e_j computes its Borda count vector $B_j = [b_j(1)\ b_j(2) \ldots b_j(N)]$, where $b_j(i) = N - \arg(r_j(k)=i)$; in other words the Borda count $b_j(i)$ for class C_i is equal to the number of classes that have a lower score than class C_i.

To combine the opinion of the M experts, the Borda count vectors are summed to form a Borda count vector for the multi-expert system:

$$B_E = [b_E(1) \quad b_E(2) \quad \ldots \quad b_E(N)] \qquad (8a)$$

$$B_E = \left[\sum_{j=1}^{M} b_j(1) \quad \sum_{j=1}^{M} b_j(2) \quad \ldots \quad \sum_{j=1}^{M} b_j(N) \right] \qquad (8b)$$

The multi-expert system classifies x by choosing the class with the highest Borda count ($E(x) = C_k$, $k = \arg\max b_E(i)$). Ties can be broken by favoring a particular expert over the others.

Evidence theory

This method of combination uses the scores generated on a training set of data to provide an estimate of the noise of each expert. The scores from the training set computed by a given expert e_j are first offset to have a mean of zero. Let the zero-mean scores be denoted as $z_j(i) = s_j(i) - \text{offset}$. The scores associated with the true class are eliminated, leaving a set of scores related to the incorrect classes, which is considered the noise of expert e_j. The standard deviation σ_j of the noise is computed and used to form a Gaussiun cumulative distribution function (CDF) G_j (mean of $2\sigma_j$, standard deviation σ_j). This CDF is used to produce the BPAs.

For a given classification sample $x \in C_k$, the mean of the resulting score vector from expert e_j is removed. Using the CDF, portions of belief are assigned to the sets $A_j(p) = \{C_i : i \in \{R_j(q) : q \leq p < N\}\}$ according to the rule:

$$m_j(A_j(p)) = G_j(z_j(r_j(p))) - G_j(z_j(r_j(p+1))) \qquad (9)$$

The remaining portions of belief are assigned to the whole set Θ. The set $A_j(p)$ contains the classes with a score equal to the p^{th} highest score or higher. The portion of belief assigned to set $A_j(p)$ is equal to the difference in CDF evaluated at the p^{th} highest score and the $(p+1)^{th}$ highest score.

Assigning the portions of belief in this manner makes the difference in plausibility between two classes proportional to the difference in their scores. The assignment is also nonlinear, which makes the difference in plausibility between two classes small if they both have very high or both have very low scores.

The BPAs from each expert are combined using Dempster's rule of combination and the plausibility is computed from the combined BPA. The multi-expert system classifies x by choosing the class with the highest plausibility ($E(x) = C_k$, $k = \arg\max Pl(C_i)$).

METHODOLOGY

Data collection

Five Canadian-English speaking male subjects, with no known speech disorders, participated in this study. A 10-word vocabulary was used, consisting of the digits "zero" to "nine". For each subject 13 sets of 40 words were constructed, with each set containing four repetitions of each word in the vocabulary. The order of the words were randomly permuted and presented to the subject one at a time, with at least one second between words to minimize coarticulatory and anticipatory effects.

During each set of words, acoustic white noise was being generated. Six different levels of noise power were used (0, 6, 9, 12, 15, and 18 dB). The absolute power of the noise was adjusted for each subject, such that the amplitude of the highest noise level (18 dB) was approximately equal to the amplitude of the acoustic speech. Of the 13 sets of words, the first, seventh, and last set used the 0 dB noise level. Each of the remaining noise levels were used twice and the order was randomly permuted among the remaining 10 sets.

Surface MES were obtained from five sites on the face (Figure 1) using Ag-AgCl Duo-Trode electrodes (Myotronics Inc.), with an Ag-AgCl RedDot electrode (3M) placed at the back of the neck providing a common ground. The MES was bandlimited between 0.1 and 500 Hz and simultaneously sampled with the acoustic signal, which was bandlimited between 0.1 and 5000 Hz. The sampling rate was 10 kHz.

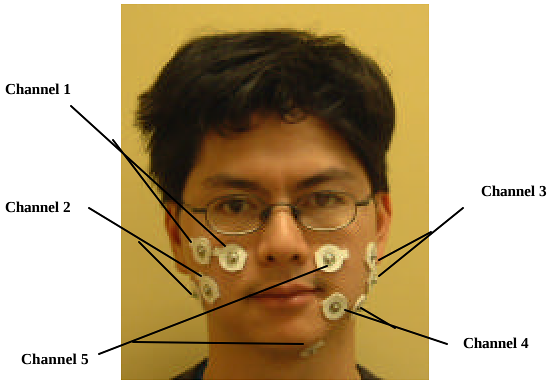

Figure 1 Electrode placement

Data were processed offline. Each word utterance was segmented using the acoustic channel as a trigger channel. The MES data were downsampled to a sampling rate of 1000 Hz and segmented into 1024 ms records. A pretrigger value of 500 ms was used, which was found to be optimal for MES ASR [4]. The acoustic data were segmented into 819.2 ms records, using a 300 ms pretrigger value.

Training and testing

The HMMs were trained using the data from the three 0 dB noise level sets (12 training examples per word) and tested on the data from the remaining noise level sets (8 test examples per word). To test at the 0 dB noise level, a leave-one-out method was used. In this method, the HMMs were trained using two of the 0 dB noise level sets and tested on the remaining 0 dB noise level set. The method was repeated to test all three 0 dB noise level sets. For the 0 dB noise level tests, there are only 8 training examples per word but a total of 12 test examples per word.

Experts

The MES expert was a 12-state, left-right hidden Markov model (HMM), with single mixture observation Gaussian densities. Overlapping observation windows of 128 ms were used, with a spacing of 16 ms. For each observation window, 13 features were computed for each MES channel: the RMS value and the first 12 mel-frequency cepstral coefficients (MFCC).

The acoustic expert was a 12-state, left-right HMM, with single mixture observation Gaussian densities. Overlapping observation windows of 25.6 ms were used, with a spacing of 12.8 ms. For each observation window, the first 12 MFCCs were used as features.

The likelihoods computed by the HMMs were normalized by the number of transition and observation probabilities, and these normalized likelihoods were considered the score vector outputs of the experts.

Combining experts

Classification results were obtained for acoustic expert (E_A) the MES expert (E_M), combining the experts using the Borda count method favoring the acoustic expert during ties (Borda$_A$), combining the experts using the Borda count method favoring the MES expert during ties (Borda$_M$), and combining the experts using evidence theory.

Table 1 Classification rates

Noise power	Acoustic	MES	Borda$_A$	Borda$_M$	Evidence
0 dB	97.7%	71.5%	91.0%	86.7%	94.8%
6 dB	92.5%	81.5%	95.3%	92.5%	93.0%
9 dB	83.8%	86.0%	93.3%	93.8%	91.0%
12 dB	36.5%	78.8%	58.0%	65.5%	80.8%
15 dB	16.8%	84.3%	39.0%	46.0%	84.0%
18 dB	11.5%	77.3%	27.8%	34.5%	77.3%

RESULTS

The classification results, averaged over the five subjects are shown in Table 1. For the acoustic expert, the classification rate is 97.7% at 0 dB but a severe decrease in classification rate can be seen as noise power is increased. The MES expert shows no trend with increasing noise power, with an average classification rate of 79.9% computed across all subjects and noise levels. Note the classification rate of the MES expert at 0 dB is slightly poorer than the rest and is probably a result of the reduced number of training examples used for this level.

The Borda count methods have classification rates equal or better than the acoustic expert in all cases except at the 0 dB level. The decrease in classification rate for the 0 dB is probably due to the poor performance of the MES expert. Although the Borda classification rates are above the classification rates of the acoustic expert, they are lower than the classification rates of the MES expert at noise levels greater or equal to 12 dB. When the acoustic expert is favored in the Borda count method, superior performance is shown at the lower noise levels. When the MES expert is favored in the Borda count method, superior performance is shown at the higher noise levels.

Combining the experts using evidence theory, classification rates of the acoustic expert are improved in all cases except the 0 dB level. The degradation in classification rate (2.9%) is less than the Borda count methods. Unlike the Borda count methods, at high noise ($\geq$ 12 dB) the performance closely follows that of the MES expert. This is the desired result because at these noise levels the acoustic expert is approaching random guessing and therefore providing no useful information; thus the highest classification rate possible should be equal to the MES expert's classification rate.

CONCLUSIONS

The method of combining the opinions of experts using evidence theory uses the differences in scores as a gauge of the reliability of an expert's opinion. As the noise power increases, the reliability of the acoustic expert accordingly decreases, while the reliability of the MES expert remains constant. Evidence theory allows a dynamic combination of the opinion of experts.

REFERENCES

1. Research and Technology Organization (North Atlantic Treaty Organization), "Alternative control technologies," RTO Technical Report 7, 1998.
2. Chan ADC, Hudgins B, Lovely DF, "Myoelectric signals in speech recognition," 26th Annual Conference of the Canadian Medical & Biological Engineering Society, Halifax, Canada, 2000.
3. Chan ADC, Englehart K, Hudgins B, Lovely DF, "Hidden Markov model classification of myoelectric signals in speech," 23rd Annual International Conference of the IEEE-EMBS, Istanbul, Turkey, 4.2.6-4, 2001.
4. Chan ADC, Englehart K, Hudgins B, Lovely DF, "Myoelectric signals to augment speech recognition," Medical & Biological Engineering & Computing, 39(4):500-504, 2001.
5. Dempster AP, "Upper and lower probabilities induced by a multivalued mapping," Annals of Mathematical Statistics, 38:325-339, 1967.
6. Shafer G, A mathematical theory of evidence, Princeton University Press, Princeton, New Jersey, 1976.
7. Black C, The theory of committees and elections, 2nd edition, Cambridge University Press, London, 1963.

THIS WORK WAS SUPPORTED BY THE DEFENSE EVALUATION RESEARCH AGENCY (DERA), UNITED KINGDOM AND THE NATURAL SCIENCES AND ENGINEERING RESEARCH COUNCIL (NSERC) OF CANADA