TWO-MICROPHONE VOICE ACTIVITY DETECTION BASED ON THE HOMOGENEITY OF THE DIRECTION OF ARRIVAL ESTIMATES

Juan E. Rubio, Kentaro Ishizuka, Hiroshi Sawada, Shoko Araki, Tomohiro Nakatani, and Masakiyo Fujimoto

NTT Communication Science Laboratories, NTT Corporation
jerm@kth.se, {ishizuka, sawada, shoko, nak, masakiyo}@cslab.kecl.ntt.co.jp

ABSTRACT

Voice Activity Detection (VAD) systems have been the object of continuous research during the last three decades. While single microphone systems cannot take advantage of certain spatial properties of speech signals, microphone array systems consisting of many elements based on beamforming techniques can be difficult to implement in reality due to cost and complexity issues. The aim of the work described in this paper was to achieve both practical feasibility and spatial discrimination ability. A new approach is developed for two-microphone VAD capable of profiting from the concentration of speech energy in time, frequency and space. The algorithm is implemented and compared with several standard VAD algorithms, such as AFE, AMR and G.729B, and other recently proposed systems, revealing promising results under real-world noise conditions. The main advantage of the proposed approach is its capacity to outperform the above methods without the need for any spatial or spectral constraints, which makes it both versatile and capable of further improvement.

Index Terms— Speech processing, Acoustic signal detection, Robustness, Direction of arrival estimation, Acoustic arrays

1. INTRODUCTION

Determining exactly when speech starts and finishes is a very complex task when the acoustic environment is filled with non-stationary noise. Unfortunately, this is the usual case when dealing with real-world applications, where there is a need for systems that can perform speech endpointing with accuracy and reliability. This is the goal of robust Voice Activity Detection (VAD), a field that is receiving considerable attention because of its relationship to, for example, speech recognition [1] and speech enhancement [2].

Recently, the use of microphone arrays for VAD purposes has been shown to be beneficial, since the spatial features of speech sounds can be used for speech/noise discrimination. Some authors suggest the use of arrays of an indefinite number of elements, based on, for instance, the calculation of the global Signal-to-Noise Ratio (SNR) using a microphone array to estimate the individual SNR for every frequency [3], or the Generalized Likelihood Ratio Test applied to multichannel input and far-field wideband sources [4]. Most of the research in the field has not addressed the problem of scaling the performance with the number of microphones. When the most important goal is high performance, we believe that microphone array techniques with many elements are the most effective, fully justifying the effort required to realize a complex design and its implementation.

However, we also think that simpler methods deserve the attention of the research community. The technique proposed in this paper takes full advantage of just two sensors, and provides a good trade-off between complexity and performance. At the same time, we have avoided beamforming approaches, because these techniques require either *a priori* knowledge of the incoming Direction of Arrival (DOA) or a certain source tracking algorithm [5], usually very sensitive to noise perturbations. Despite its simplicity, our system uses both spatial and spectral features of the source, unlike the majority of existing algorithms, which focus on either the spectral or spatial [6] characteristics.

We have assumed a certain spatial characterization for both speech and noise sources. As regards speech, the organs involved in its production happen to be very small, especially in relation to the distance to and between the microphones. This means that a speech source is spatially perceived as an object that occupies not more than a few degrees. At the same time, it is not usually very far from the microphones and is pointed at them, which creates a direct dominant propagation path and, in spite of significant room reverberation, a clear DOA estimate can be obtained. Regarding noise sources, some are significantly bigger in size (vibrating walls, engines, big crowds...) and their DOA is highly unlocalized. However, even when their size is comparable to that of a speech source, they are usually located further away, which results in multipath propagation and DOA spread. These are the key differences between speech and noise sources that we try to exploit in our system.

In short, the proposed technique is based on a new decision measure that represents the degree to which sound energy concentrates in space for a certain time-frequency region. This measure is referred to as DOA homogeneity, and it is defined based on the entropy of the DOA estimations, determined from the 2-channel observed signal. If we adopt the above assumptions, the proposed measure provides a higher value for speech regions than for noise regions, a fact that makes it possible to distinguish and classify them. These DOA homogeneity values, which are represented as a two-dimensional map, finally become the input for the statistical binary classifier proposed as a VAD system in [8].

The tests employ Receiver Operating Characteristic (ROC) curves to compare our proposed method with the conventional Statistical VAD (SVAD) [8], as well as with other widely deployed standards published by ETSI and ITU and recently developed algorithms, namely Long-Term Spectral Divergence

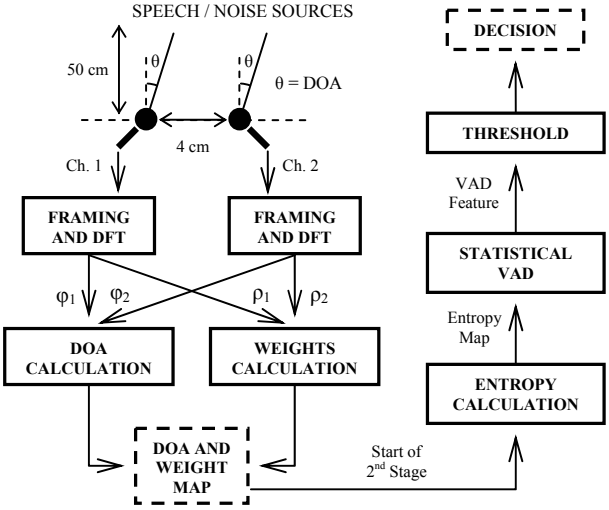

Figure 1. *Schematic of the proposed algorithm.*

(LTSD) [9], First-Order Differential Microphone (FODM) [1] and Magnitude Squared Coherence (MSC) [2]. We also investigate the robustness of our algorithm with respect to changes in the noise environment and the Signal-to-Noise Ratio (SNR) to enable us to draw more general conclusions about its performance.

2. ALGORITHM DESCRIPTION

2.1. DOA Estimation and Weight Calculation

The proposed algorithm, shown in Fig. 1, consists of two stages. First, both channels of the signal are divided into frames of length L with a certain amount of overlapping. Every frame is multiplied by a Hanning window, followed by a DFT operation. Given m as the microphone index, the observations $x_m(f,t)$ are expressed as in (1), resulting in the DFT magnitudes $\rho_m(f,t)$ and phases $\varphi_m(f,t)$, which depend on frequency f and time t.

$$x_m(f,t) = \rho_m(f,t)\cdot e^{j\varphi_m(f,t)} \tag{1}$$

The obtained DFT phases $\varphi_1(f)$ and $\varphi_2(f)$ are used to estimate the DOA for each frequency bin [7]. Taking account of geometric considerations and a far-field assumption, we express the DOA $\theta(f,t)$ as in equation (2), where v_s corresponds to the speed of sound and d is the distance between the microphones.

$$\theta(f,t) = \frac{\arcsin\left(v_s\cdot(\varphi_1(f,t)-\varphi_2(f,t))\right)}{2\cdot\pi\cdot d\cdot f} \tag{2}$$

At the same time, the DFT magnitudes are used to weigh their corresponding DOA estimates. Based on the principle that frequencies with high SNR produce more reliable estimates, the weights consist of the sum of the powers from both channels, as in equation (3). Additionally, certain weights are set at zero when, due to additive noise effects, the inverse sine function receives a phase difference outside the interval [–1, 1] as an argument.

$$\begin{aligned}\theta(f,t) &\leftarrow W(f,t)\cdot\theta(f,t)\\ W(f,t) &= \rho_1^2(f,t)+\rho_2^2(f,t)\end{aligned} \tag{3}$$

2.2. Entropy calculation

Our goal is to find small time-frequency regions where the DOA is homogeneous. These analysis regions are small rectangular grids, with a time width N and frequency height M, as in expression (4).

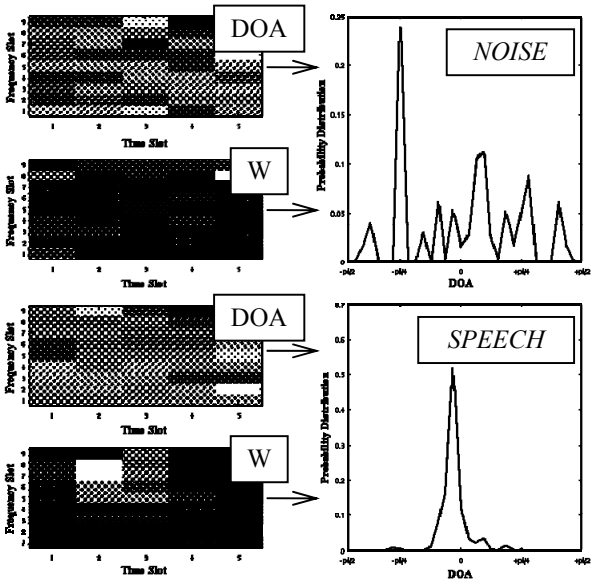

Figure 2. *Examples of noise (top) and speech (bottom) regions, showing the DOA estimates (left, top), power-based weights W (left, bottom) and resulting histogram (right) for each case.*

$$\begin{aligned}&\mathrm{GRID}(t_i,f_j) = \{(t_{i+P},f_{j+Q})\}\\ &\forall P\in\left\{-\frac{N-1}{2},\cdots,+\frac{N-1}{2}\right\};\ \forall Q\in\left\{-\frac{M-1}{2},\cdots,+\frac{M-1}{2}\right\}\end{aligned} \tag{4}$$

To calculate the homogeneity of these $N \times M$ DOA estimates, a histogram is built to approximate the real distribution. The continuous DOA space [–π/2, +π/2] is quantized into B bins, while the contribution of each individual estimate to the histogram varies depending on its corresponding power-based weight, as shown in Fig. 2. The entropy of this distribution is calculated as in expression (5). Here, $p_b(f,t)$ is the probability of DOA bin b, calculated for a grid centered at frequency f and time t. The term $\bar{\theta}(f,t)$ corresponds to the quantized DOA estimations and θ_b are the quantization bins.

$$\begin{aligned}H(f,t) &= -\sum_{b=1}^{B} p_b(f,t)\cdot\log_2(p_b(f,t))\\ p_b(f,t) &= p\left(\bar{\theta}(f,t)=\theta_b\right)\end{aligned} \tag{5}$$

We decided to use this entropy feature because, when we already have a low score due to speech presence and certain interference appears, the outcome remains almost unaffected. Other features, such as the standard deviation, do not have this advantage. Finally, DOA homogeneity $\Delta(f,t)$ is determined by inverting and normalizing the entropy, as shown in equation (6). The maximum entropy value H_{MAX} depends on the number of bins B into which the DOA is quantized and is calculated as the entropy of a completely flat distribution. Substituting equal probability values into equation (5), we obtain $H_{MAX} = \log_2(B)$. The result $\Delta(f,t)$ is a normalized time-frequency map of the inverted DOA entropy values, which range from 0 (noise, complete randomness), to 1 (speech, total organization).

$$\Delta(f,t) = (1 - H(f,t))/H_{MAX} \quad (6)$$

$\Delta(f,t)$ follows a normal distribution whose variance becomes larger in the presence of speech. As in [8], we apply thresholding to the log-likelihood ratio of speech to non-speech probabilities, assuming normal distributions. Let us explore the features of these entropy maps from two different points of view:

1) *Time-Frequency domain*: Other methods make direct use of the concentration of speech power around harmonics, but our system uses this information indirectly, modifying the DOA estimation and making it more reliable. In addition, SVAD introduces discrimination ability for the frequency patterns that appear in the entropy maps. Grid processing also introduces a certain smoothing, which reduces the effect of the noise and limits the time resolution.

2) *Spatial domain*: The main advantage of our method relates to this domain. Considering the assumptions we have made, our algorithm is capable of distinguishing physically small sources that are relatively close to the microphones, and which therefore have a clear propagation path (speech) from other more unlocalized sources that lie further away or are physically bigger (noise). The robustness of our algorithm in the presence of non-stationary noise compensates for its weakness with certain kinds of interference, validating the assumptions we made.

3. EVALUATION

3.1. Experimental setup

To test the system, a total of 14 men and women spoke a sentence in Japanese, which was recorded using two omni-directional microphones, placed in a line perpendicular to the incoming sound and 4 cm apart. We maintained the distance to the microphones at one order of magnitude greater, to allow us to assume the far-field hypothesis. The incoming DOA θ was always measured as the angle of deviation from a line orthogonal to the microphones.

The background noise was recorded in Tokyo, Japan using a similar configuration, surrounded by different non-stationary real environments: airport, train platform, train ticket gate, restaurant, subway platform, subway ticket gate and street.

3.2. Description of tests

The sampling frequency was decided experimentally and set at 8 kHz, the frame length L was 32 ms with a 50% overlap, the DFT operations were computed using 256 points, the width and height of the entropy grids, N and M, were 9 and 5, respectively, and the number of DOA quantization bins B was set at 32.

For every combination of the 7 noise environments and 14 speech utterances, 4 different SNR values were considered: –5, 0, 5 and 10 dB. In addition, 2 different noise realizations were used for every combination in order to obtain a more reliable average. This meant that 784 different noisy speech utterances were tested. The SNR was defined as the ratio between speech and noise energy, calculated from the first to the last speech sample in the utterance. Every test output a continuous feature, to which we applied simple binary thresholding. The accuracy of the resulting VAD decision was obtained by comparing it with the manual labels, and expressed as the relationship between the false acceptance rate, namely the percentage of incorrectly detected speech frames, which are actually noise, and the false rejection rate, namely the percentage of incorrectly detected noise frames, which are actually speech.

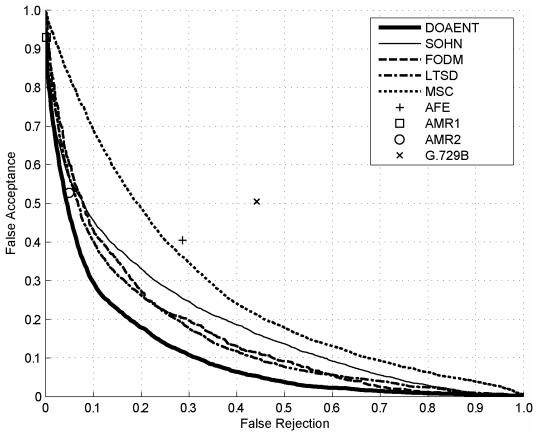

Figure 3. *ROC results at 0 dB SNR: our method (DOAENT), Sohn's Statistical VAD, FODM, LTSD and MSC algorithms, and the standards AFE, AMR1/2 and G.729B. Airport noise environment.*

For each environment and SNR, the obtained data were plotted as ROC curves. The rest of the compared algorithms, LTSD, FODM, MSC and SVAD, received the same treatment. We chose these methods for different reasons; SVAD is directly related to our system, FODM and MSC focus specifically on two-microphone systems, and LTSD is a fairly simple single channel algorithm with good performance. The standards ETSI Advanced Front-End (AFE) [10], ETSI Adaptive Multi-Rate (AMR1 and AMR2) [11] and ITU-T G.729B [12], were plotted as single points in the ROC space due to their optimal threshold specification. To make a fairer comparison, single-channel methods were provided with a beamformed signal, using a delay-and-sum beamformer aimed at 0 degrees (similar spatial constraint as FODM), whose approximate SNR gain is 3 dB. We also carried out tests using only one of the channels, although the results were never better.

3.3. Results

To provide an illustrative example, we show detailed results for 0 dB SNR and noise recorded at an airport. However, more general results in terms of both SNR and noise environment are given in numerical form. As seen in Fig. 3, the best performing ROC curve belongs to our DOA-Entropy algorithm. In terms of SNR gain, the second best system (FODM) performs at an estimated 4-5 dB below ours. Moreover, it is interesting to notice the great improvement achieved simply by changing the original SVAD input to our DOA-Entropy map.

With respect to changes in SNR, Equal Error Rates (EER) and working points are shown in Tables 1 and 2, respectively. EER consist of the points in the ROC space where the false acceptance and false rejection rates are equal. Qualitatively, the order of the best performing algorithms was basically the same in all the cases, namely DOA-Entropy in first position and FODM second, while the other algorithms performed worse on average (see Table 1). More specifically, at –5 dB SNR FODM performed slightly better, while at 0 dB FODM and DOA-Entropy performed equally well, and at 5 and 10 dB our method was superior with an improvement margin of over 10%. This is mainly because the assumption made for the weighting

Table 1. *Equal Error Rates, averaged for seven different noise environments. The rows show the tested SNR values; the columns show the different VAD algorithms.*

	DOA	*FODM*	*SOHN*	*LTSD*	*MSC*
Mean 10 dB	8.1%	9.3%	16.7%	9.6%	12.6%
Relative 10 dB	*Ref.*	*14.8%*	*106.0%*	*18.5%*	*55.5%*
Mean 5 dB	12.1%	13.6%	20.9%	15.7%	17.8%
Relative 5 dB	*Ref.*	*12.4%*	*72.7%*	*29.7%*	*47.1%*
Mean 0 dB	19.5%	19.6%	25.6%	25.9%	24.8%
Relative 0 dB	*Ref.*	*0.5%*	*31.3%*	*32.8%*	*27.2%*
Mean -5 dB	28.9%	27.2%	33.4%	35.4%	33.7%
Relative -5 dB	*Ref.*	*-5.8%*	*15.6%*	*22.5%*	*16.6%*

Table 2. *Standards' working points, averaged for the different noise environments. The first percentage corresponds to the false rejection rate and the second to the false acceptance rate.*

	AFE	*G.729B*	*AMR2*
Mean 10 dB	4.2%, 45.1%	16.3%, 53.1%	0.3%, 71.0%
Mean 5 dB	12.0%, 41.3%	22.1%, 53.9%	0.5%, 70.4%
Mean 0 dB	22.1%, 43.6%	36.3%, 54.1%	3.6%, 66.4%
Mean -5 dB	33.1%, 42.3%	57.9%, 53.0%	23.0%, 56.3%

procedure holds better with a positive SNR. When there is too much noise, high power at a certain frequency is less likely to be caused by speech, so that noisy DOA estimations contribute significantly to the histogram estimation, making it 'whiter'.

The working points (false rejection and false acceptance rates) of the different standards, shown in Table 2, were always above the ROC curve of our algorithm, which indicates inferior performance. These results are approximately in accordance with those reported in [9] for the AURORA-2 database [13].

To illustrate the advantage of the absence of spatial constraints for our algorithm, in Fig. 4. we show the impact of modifying the speech's incoming DOA. We took the only noise environment ('restaurant') where FODM clearly outperforms our method and set the SNR at 0 dB. As we can see, using a DOA of just 32º produces a radical change, since the ROC curve for FODM becomes useless. However, the DOA-Entropy performance does not remain unaffected because of the variation in the resolution at different angles and the higher impact of noise. Nevertheless, this performance degradation is much less harmful than with FODM. Therefore, our algorithm is very useful with moving speakers, where there are no spatial constraints.

4. CONCLUSION

The proposed algorithm outperforms other recent systems and standards, intended for both dual and single microphone operation, without the need for certain constraints and in adverse noise environments with many highly localized interferences.

Additional processing in the spectral, spatial or time domains should produce further improvements. One especially interesting problem is how to suppress other undesired interfering sources whose DOA estimations are homogenous, resembling those for the target speech. When the DOA is significantly different, it is possible to use either a more sophisticated weighting procedure or to manipulate the DOA estimations directly.

Although this is inherently a two-microphone technique, it is possible to integrate information from other microphones by creating different pairs with independent estimations, provided

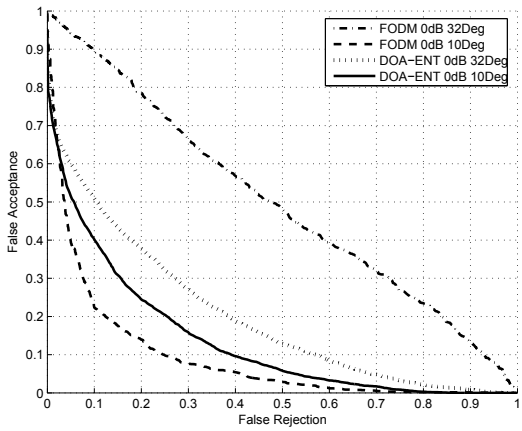

Figure 4. *ROC curves using DOA-Entropy and FODM methods for different angles: 0 and 32 degrees. Restaurant noise environment.*

we have a good geometrical design. In this way, we would be able to compensate for low precision at side angles, and to detect more than one spatial dimension for speech source localization.

REFERENCES

[1] A. Álvarez, P. Gómez, V. Nieto, R. Martínez, and V. Rodellar, "Application of a First-Order Differential Microphone for Efficient Voice Activity Detection in a Car Platform", *Proc. Interspeech*, pp. 2669-2672, 2005.

[2] R. Le Bouquin-Jeannès and G. Faucon, "Study of a Voice Activity Detector and its Influence on a Noise Reduction System," *Speech Communication*, Vol. 16, pp. 245–254, 1995.

[3] J. Chen and W. Ser, "Speech Detection using Microphone Array", *Electronic Letters*, Vol. 36, No. 2, pp. 181-182, Jan. 2000.

[4] I. Potamitis, "Estimation of Speech Presence Probability in the field of Microphone Array", *IEEE Signal Processing Letters*, Vol. 11, No. 12, pp. 956-959, December 2004.

[5] D. Van Compernolle and S. Van Gerven, "Beamforming with Microphone Arrays", *Applications of Digital Signal Processing to Telecommunications*, pp. 107-131, E. U., COST 229, 1995.

[6] T. Pirinen and A. Visa, "Signal Independent Wideband Activity Detection Features for Microphone Arrays", *Proc. of ICASSP 2006*, Vol. IV, pp. 1109-1112.

[7] J. Benesty, S. Makino, and J. Chen, *Speech Enhancement*, Springer, The Netherlands, pp. 308-310, 2005.

[8] J. Sohn, N. Kim, and W. Sung, "A Statistical Model-Based Voice Activity Detector", *IEEE Signal Processing Letters*, Vol. 16, No. 1, pp. 1-3, 1999.

[9] J. Ramírez, J. C. Segura, C. Benítez, Á. de la Torre, and A. Rubio, "Efficient Voice Activity Detection Algorithms using Long-Term Speech Information", *Speech Communication*, Vol. 42, pp. 271-287, 2004.

[10] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-End Feature Extraction Algorithm; Compression Algorithms", *ETSI ES 202 050*, v1.1.4, 2005.

[11] ETSI, "Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels", *ETSI EN 301 708*, 1999.

[12] ITU, "A Silence Compression Scheme for Use with G.729 Optimized for V.70 Digital Simultaneous Voice and Data Applications", *ITU-T Recommendation G.729-Annex B,* 1996.

[13] H. G. Hirsch and D. Pearce, "The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions", *Proc. of Interspeech*, vol. 1, pp. 341–344, 2000.