IDIAP RESEARCH REPORT

Observations on Multi-Band Asynchrony in Distant Speech Recordings

Guillaume Lathoud [a,b]

IDIAP–RR 06-74

December 2006

a IDIAP Research Institute, CH-1920 Martigny, Switzerland
b Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 06-74

Observations on Multi-Band Asynchrony in Distant Speech Recordings

Guillaume Lathoud

December 2006

Abstract. Whenever the speech signal is captured by a microphone distant from the user, the acoustic response of the room introduces significant distortions. To remove these distortions from the signal, solutions exist that greatly improve the ASR performance (what was said?), such as dereverberation or beamforming. It may seem natural to apply those signal-level methods in the context of speaker clustering (who spoke when?) with distant microphones, for example when annotating a meeting recording for enhanced browsing experience. Unfortunately, on a corpus of real meeting recordings, it appeared that neither dereverberation nor beamforming gave *any* improvement on the speaker clustering task. The present technical report constitutes a first attempt to explain this failure, through a cross-correlation analysis between close-talking and distant microphone signals. The various frequency bands of the speech spectrum appear to become desynchronized when the speaker is 1 or 2 meters away from the microphone. Further directions of research are suggested to model this desynchronization.

1 Introduction

The information flow from humans to computers is most often limited by keyboard usage because the information flow is usually much less than in speech, and because keyboard usage hampers interactions between humans. Removing this bottleneck would transform computers into truly helpful assistants. One possible direction is "distant speech processing", where the information flow comes from the speech signal acquired by a microphone untethered to, and distant from the user. The acoustic response of the room introduces significant distortions into the captured speech signal. To remove these distortions, solutions exist that greatly improve the Automatic Speech Recognition (ASR) performance (what was said?), such as dereverberation [1] or beamforming. It would be natural to apply those signal-level methods in the context of speaker clustering (who spoke when?), for example when annotating a meeting recording for enhanced browsing experience. Unfortunately, on a corpus of real meeting recordings [2], it appeared that neither dereverberation nor beamforming gave *any* improvement on the speaker clustering task [3, Section 4.3]. The present technical report constitutes a first attempt to explain this failure, through a cross-correlation analysis between close-talking and distant microphone signals. The various frequency bands of the speech spectrum appear to become desynchronized when the speaker is 1 or 2 meters away from the microphone. Further directions of research are suggested to model this desynchronization.

The rest of this paper is organized as follows: Section 2 summarizes a previous speaker clustering experiment [3], that used distant microphones only. Section 3 provides a cross-correlation analysis between a close-talking and a distant microphones. It appears that multiple frequency bands can be desynchronized, which Section 4 attempts to explain. Section 5 concludes.

2 Summary of the Speaker Clustering Experiment

In an experiment described in details in [3], the task was unsupervised speaker clustering on meeting recordings, using distant microphones only.

- **The task** of unsupervised speaker clustering [4, 5] means that we are trying to estimate not only the precise speech/silence segmentation of time, but also the correct speaker *identity* for each speech segment, as well as the correct number of speakers. In the experiments [3], speaker clustering is applied to each meeting separately. No enrollment data is available, therefore speaker identity is defined as a numeric tag.

- **The meeting recordings** were taken from the M4 Corpus [2]. 18 meetings were used, where each meeting includes 4 speakers seated around a table, and lasts about 5 minutes. *Sometimes a speaker stands up and moves to the presentation screen or to the whiteboard* (red, dashed rectangle in Fig. 1). A very precise ground-truth speech/silence segmentation was provided by a human annotators.

- **The signals** used for the speaker clustering experiments in [3] were obtained from a 8-microphone, 10-cm radius circular array placed on the table, in the middle (red, continuous ellipse in Fig. 1).

In the results of the speaker clustering experiments [3], speech from a given speaker would correctly be grouped into a single speaker cluster, as long as the speaker was seated. This was also verified when running speaker clustering on a concatenation of three meetings, with the same speaker at different seats. However, whenever a speaker would stand up and move further away from the array, (as visible for example on the rightmost red, dashed rectangle in Fig. 1), his/her speech would *systematically* become splitted into two clusters:

- One cluster when the speaker was close to the array (i.e. seated),

- Another cluster when the speaker was far from the array (i.e. standing at the presentation screen or the wideboard).

Figure 1: Snapshots of the 3 cameras, from one meeting of the M4 Corpus [2]. Sometimes a speaker (red, dashed rectangle) stands up and goes to the presentation screen, thus moving further away from the microphone array.

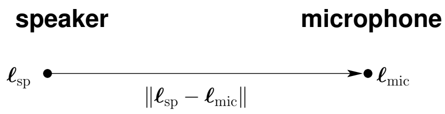


Figure 2: Point source model. $\boldsymbol{\ell}_{\mathrm{sp}} \in \mathbb{R}^3$ and $\boldsymbol{\ell}_{\mathrm{mic}} \in \mathbb{R}^3$ are spatial locations. The two signals $x_{\mathrm{sp}}(t)$ and $x_{\mathrm{mic}}(t)$ only differ by a pure delay: $x_{\mathrm{mic}}(t) \quad \propto \quad x_{\mathrm{sp}}\left(t - \frac{\|\boldsymbol{\ell}_{\mathrm{sp}} - \boldsymbol{\ell}_{\mathrm{mic}}\|}{c}\right),$ where c is the speed of sound in the air, $\|\cdot\|$ is the Euclidean norm, and $\propto$ means "is proportional to".

We tested low-level signal processing methods that have proved useful to improve MFCC-based (ASR) results in such cases, hoping that they would also improve the speaker clustering results. The tested techniques included delay-sum beamforming, dereverberation through long-term log-magnitude mean removal [1] and noise removal through short-term spectral subtraction [6]. Unfortunately, all these techniques resulted in *absolutely no performance improvement*, with respect to the above-described issue. Therefore, it seems that feature variability between close and distant locations is a bottleneck to speaker clustering with distant microphones, which suggests some basic research. Section 3 provides an observation that characterizes the distance-dependent variability.

3 Observation: Cross-Correlation Analysis

This section reports cross-correlation experiments, where we compare the signal captured by a lapel microphone near the mouth of a speaker, with the signal captured by a distant microphone. Section 3.1 briefly reminds the point source model, then Section 3.2 presents observations on real signals. The code and data used in the experiments presented in this section are fully available at: `http://mmm.idiap.ch/Lathoud/2006-distant-speaker`

3.1 Point Source Model in a Free Field

If we model the speaker's mouth as a point source in a free field environment (no reverberation), and assume the air to be an homogeneous medium (constant speed of sound), then the two signals only differ by a pure delay $\frac{\|\boldsymbol{\ell}_{\mathrm{sp}} - \boldsymbol{\ell}_{\mathrm{mic}}\|}{c}$, as described in Fig. 2. With this model, the key point is that *the medium is non-dispersive*, which means that the Time Of Flight (TOF) is the same for all frequencies

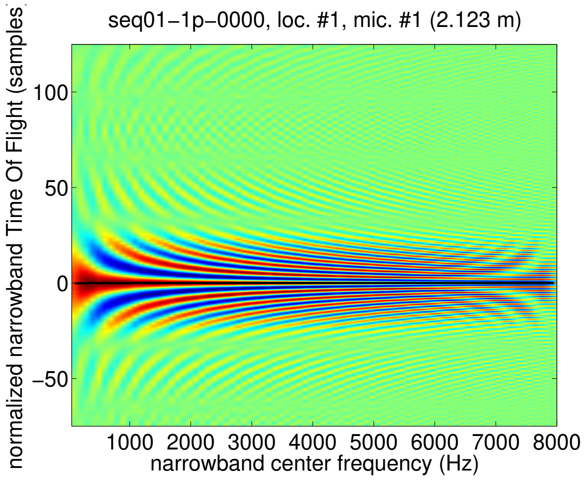


Figure 3: Narrowband theoretical cross-correlation, that is between $y_{\mathrm{mic}}(t)$ and $y_{\mathrm{sp}}(t)$ (highs in red, lows in blue, the black line marks the peak). "Normalized" means that we have subtracted $\arg\max_{\tau} g_{x_{\mathrm{mic}},x_{\mathrm{sp}}}(\tau)$, the TOF for which the *fullband* cross-correlation is maximum.

of the signal. This is a well-known characteristic of acoustic waves in air. The delay $\frac{\|\boldsymbol{\ell}_{\mathrm{sp}}-\boldsymbol{\ell}_{\mathrm{mic}}\|}{c}$ is the theoretical TOF of the acoustic wave from the mouth to the microphone.

Let us now consider the fullband cross-correlation:

$$g_{x_{\mathrm{mic}},x_{\mathrm{sp}}}(\tau) \stackrel{\mathrm{def}}{=} [x_{\mathrm{mic}}(t) \otimes x_{\mathrm{sp}}(-t)](\tau) \tag{1}$$

where τ is the TOF, and $\otimes$ designates the convolution operator. With the model described in Fig. 2, and similarly to [7, equation (9)]:

$$g_{x_{\mathrm{mic}},x_{\mathrm{sp}}}(\tau) \propto \left[g_{x_{\mathrm{mic}},x_{\mathrm{mic}}}(t) \otimes \delta\left(t - \frac{\|\boldsymbol{\ell}_{\mathrm{sp}}-\boldsymbol{\ell}_{\mathrm{mic}}\|}{c}\right)\right](\tau) \tag{2}$$

which will be maximum for the TOF $\tau = \frac{\|\boldsymbol{\ell}_{\mathrm{sp}}-\boldsymbol{\ell}_{\mathrm{mic}}\|}{c}$ (zero on the Y axis in Fig. 3).

Let us now look at narrowband signals:

$$y_{\mathrm{mic}}(t) \stackrel{\mathrm{def}}{=} [h \otimes x_{\mathrm{mic}}](t) \tag{3}$$

$$y_{\mathrm{sp}}(t) \stackrel{\mathrm{def}}{=} [h \otimes x_{\mathrm{sp}}](t) \tag{4}$$

where $h(t)$ is the impulse response of a bandpass filter. Using frequency-domain quantities (not shown here), the narrowband cross-correlation $g_{y_{\mathrm{mic}},y_{\mathrm{sp}}}$ can be expressed as:

$$g_{y_{\mathrm{mic}},y_{\mathrm{sp}}}(\tau) = \left[g_{h,h}(t) \otimes g_{x_{\mathrm{mic}},x_{\mathrm{sp}}}(t)\right](\tau) \tag{5}$$

which, using (2), becomes:

$$g_{y_{\mathrm{mic}},y_{\mathrm{sp}}}(\tau) \propto \left\{g_{h,h}(t) \otimes \left[g_{x_{\mathrm{mic}},x_{\mathrm{mic}}}(t) \otimes \delta\left(t - \frac{\|\boldsymbol{\ell}_{\mathrm{sp}}-\boldsymbol{\ell}_{\mathrm{mic}}\|}{c}\right)\right]\right\}(\tau) \tag{6}$$

$$\propto \left\{[g_{h,h}(t) \otimes g_{x_{\mathrm{mic}},x_{\mathrm{mic}}}(t)] \otimes \delta\left(t - \frac{\|\boldsymbol{\ell}_{\mathrm{sp}}-\boldsymbol{\ell}_{\mathrm{mic}}\|}{c}\right)\right\}(\tau) \tag{7}$$

which will be maximum (black line in Fig. 3) for the TOF $\tau = \frac{\|\boldsymbol{\ell}_{\mathrm{sp}}-\boldsymbol{\ell}_{\mathrm{mic}}\|}{c}$ (zero on the Y axis in Fig. 3).

3.2 Observation on Real Signals

We used a recording with a single speaker at 16 different locations around the table: standing as well as seated, close as well as far from the 16 table microphones. This is recording `seq01` from the AV16.3 Corpus [8], which was made in the same room as the M4 Corpus used in Section 2, with almost no change in the furniture. All signals were sampled at 16 kHz.

We used the lapel microphone as an estimate of the "emitted" speech signal, and one of the microphones on the table as a "distant microphone". The exact same analysis as in Section 3.1 was conducted: fullband and narrowband cross-correlation, using the GCC-PHAT estimate [7]. Time frames where 16 ms long, thus guaranteeing the stationarity of the signals (we obtained very similar results for 32 ms frames). For each of 80 narrowbands covering the interval from 50 Hz to 8 kHz, the complex frequency domain cross-correlation was averaged across all speech time frames spent by a speaker at a given location (about 6 to 8 seconds total, for each speaker location), before applying the PHAT normalization [7].

Two results are shown in Fig. 4. In the small distance case (Fig. 4a), the result is in accordance with the theory (Fig. 3). This is not the case when the speaker is further away (Fig. 4b), where **the estimated TOF becomes frequency-dependent**. We repeated this cross-correlation analysis for all 16 speaker locations and 16 table microphones, and obtained a very consistent trend, as illustrated by Fig. 5a. We also repeated all experiments on recording `seq03` (different subject)[1], obtaining extremely similar trends, as summarized by Fig. 5b.

We looked at all individual cross-correlation results (such as Figs. 4a and 4b), and observed that **the spectrum tends to be splitted into a few bands**. Within each band, the estimated TOF (black line in Fig. 4) does not vary much, and always in a continuous manner. However, **at the boundary between two bands, the estimated TOF has a sharp discontinuity or "jump"**. We'll call this phenomenon "multi-band asynchrony". Figs. 5a and 5b indicate that multi-band asynchrony increases when the speaker distance increases.

4 Interpretation

Section 3 showed that the time-averaged narrowband cross-correlation features "multi-band asynchrony", where the spectrum is splitted in a few bands, with a stable value of estimated TOF, in each band. One can distinguish two cases:

1. In a band, the estimated TOF is close to the theoretical TOF.
2. In a band, the estimated TOF is larger than the theoretical TOF.

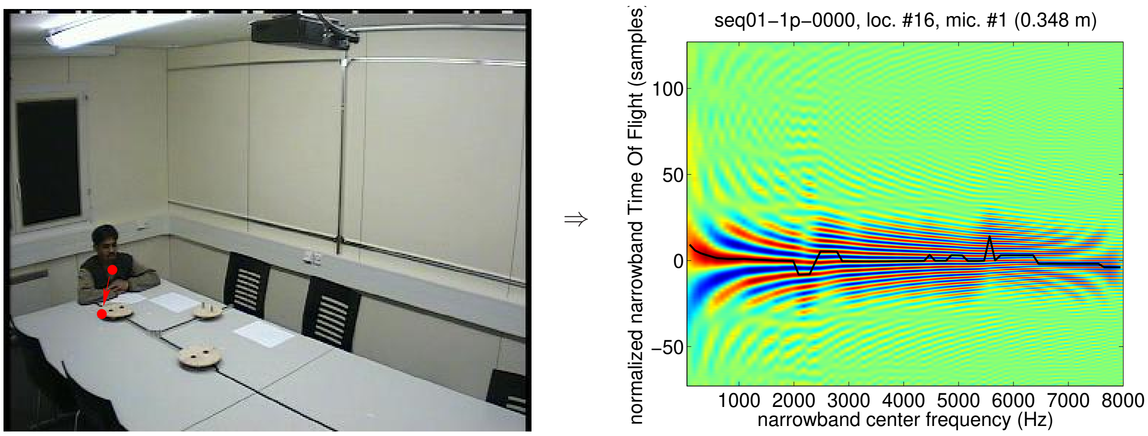
In the first case, the acoustic wave is likely to have followed the direct path. In the second case, the acoustic wave is likely to have followed an indirect path. Based on this interpretation, it would appear that *the dominant path followed by the acoustic wave depends on the frequency.*

4.1 Possible Causes

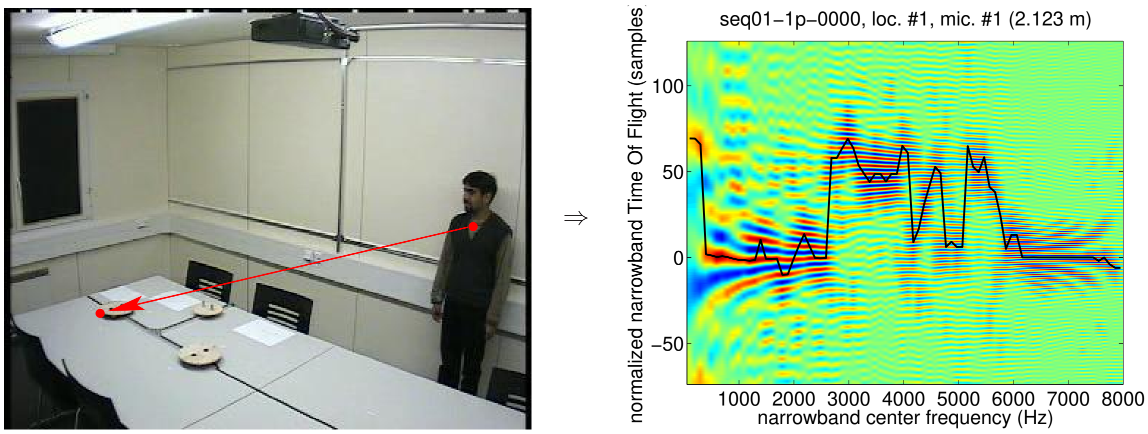
Accepting this interpretation, two possible (non-exclusive) causes can be hypothesized.

First, a frequency-dependent acoustic path could seem analogous to observations made on musical instruments [9]. However, the study in [9] considered steady, stationary musical sounds, whereas the observations reported in Section 3 were obtained by averaging over many time frames of an essentially non-stationary signal. Intuitively, one can contrast the *instantaneous, changing* mouth shape and the *long-term, stable* multi-band asynchrony observed at each speaker location. Indeed, the characteristics of human speech radiation strongly depends on the particular type of phoneme pronounced at a given time [10]. It is not clear whether, and how, these instantaneous mouth characteristics impact on the long-term statistics observed here.

[1]Recording `seq02` could not be used because it does not include a lapel microphone.

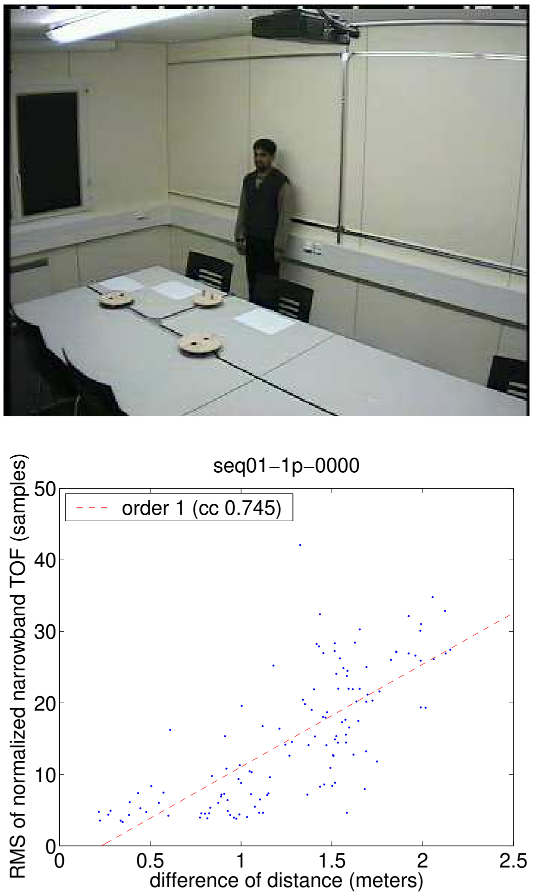


(a) Speaker close to the microphone

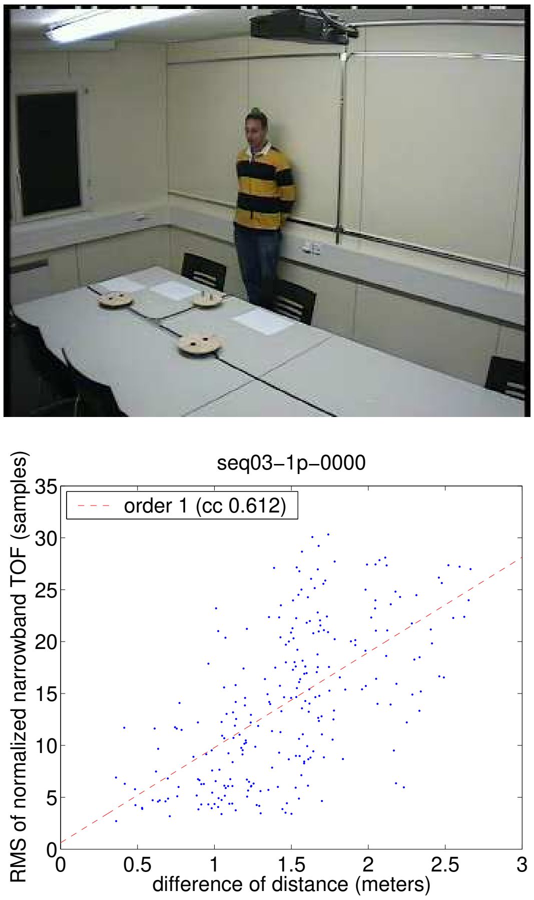


(b) Speaker far from the microphone

Figure 4: `seq01` from the AV16.3 Corpus. Cross-correlation analysis between the lapel microphone (red dot near the throat) and a distant microphone (red dot pointed by the arrow). The legend is the same as in Fig. 3.

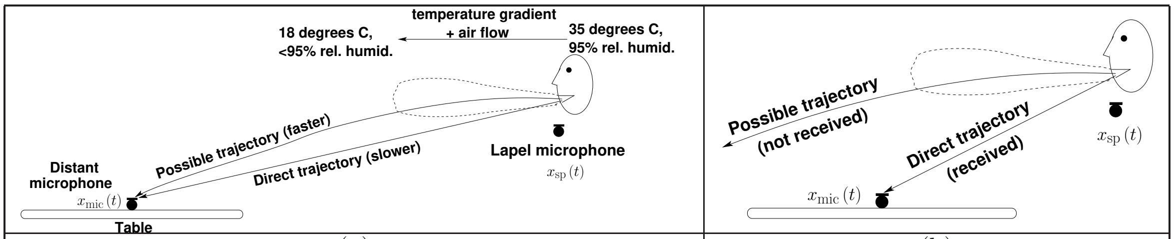


(a) seq01



(b) seq03

Figure 5: Two single-speaker recordings seq01 and seq03, from the AV16.3 Corpus. Top row: for each recording, screenshot of one of the 16 speaker locations. Bottom row: for each recording, summary of the cross-correlation analysis for all 16 speaker locations and all 16 table microphones. "RMS" stands for Root Mean Square TOF across all narrowbands (RMS of the black line in Fig. 4). Each dot correspond to one pair (speaker location, table microphone). The dashed line depicts the linear regression result, and "cc" stands for correlation coefficient between X values and Y values.



(a) (b)

Figure 6: Possible modification of the trajectory of a wave, depending on the speaker's distance to the microphone. Room reverberations are not taken into account in these figures.

Alternatively, one can remember that in Section 3.1, we assumed a non-dispersive *medium*, based on an underlying assumption of homogeneity of the medium. The latter assumption implies constant air temperature and constant humidity, over the room space. However, the medium is certainly not homogeneous, because the speaker is emitting a jet of hot, humid air (about 35^oC and 95 % relative humidity [11]), in the middle of drier, colder air. This fact reminds studies conducted on larger turbulent systems such as seismic infrasonic waves through the troposphere turbulences [12]. These studies led to *dispersive* models of the global transmission channel, where different frequencies arrive at different times. The dispersive assumption would be linked to the presence of turbulence(s), which does not contradict the well known, non-dispersive characteristics of homegeneous air. As opposed to the other possible cause mentioned above, at each speaker location the exhaled jet of hot air can be considered as steady, at least in `seq01` and `seq03`, because speakers were not moving much while speaking. It is thus reasonable to assume that, *if* the exhaled jet of hot air has an impact on the acoustic transmission channel, then this impact should be reflected by long-term statistics such as the average cross-correlation used in Section 3.

Let us now assume that the exhaled jet of hot air indeed modifies the acoustic transmission channel. We can think this jet as a "lossy waveguide", where lossy means that its boundaries let pass some of the acoustic power. The sharp discontinuities in terms of TOF, observed in Section 3.2, may then be linked to, and possibly explained by, on one hand, the particular dimensions and associated resonance modes of this lossy waveguide, and on the other hand, the large range of wavelengths that compose speech. The two cases mentioned at the beginning of the present section would then be explained as follows. In both `seq01` and `seq03`, the speaker is exhaling in a somewhat horizontal direction, above the table plane. For a distant location (lower elevation), the gradient of temperature may lead to a curved trajectory of the dominant acoustic wave received by the distant microphone (signal $x_{\mathrm{mic}}(t)$), as illustrated in Fig. 6a. This may be less the case when the speaker is closer to the array (higher elevation, as in Fig. 6b). Sections 4.3 and 4.2 verify two consequences of this hypothesis. Data and code for all experiments reported in Sesctions 4.2 and 4.3 are fully available at: `http://mmm.idiap.ch/Lathoud/2006-distant-speaker`

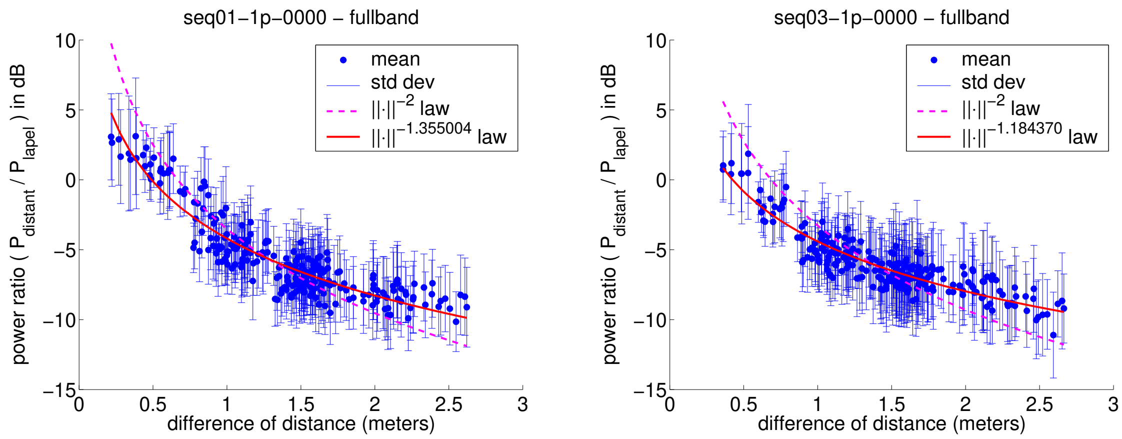
4.2 Observation: Power Decay

If we assume that a spherical wave is emitted at $\boldsymbol{\ell}_{\mathrm{sp}} \in \mathbb{R}^3$, and travels in a free field, then the "inverse square law" should be verified, where at a given spatial location $\boldsymbol{\ell}_{\mathrm{mic}} \in \mathbb{R}^3$, the received power (square of the amplitude) is proportional to $\|\boldsymbol{\ell}_{\mathrm{sp}} - \boldsymbol{\ell}_{\mathrm{mic}}\|^{-2}$. On the other hand, if we now assume a "waveguide" of hot air (Fig. 6), then the acoustic wave should be less dispersed that in the spherical wave case, therefore the transmitted power should decrease slower than in the inverse square law $\|\boldsymbol{\ell}_{\mathrm{sp}} - \boldsymbol{\ell}_{\mathrm{mic}}\|^{-2}$. Fig. 7 confirms this expectation. The continuous line represents the Minimum Mean Square Estimate (MMSE) of the exponent in dB domain, obtained by initializing with the square law (initial exponent value =-2), and then applying a few steps of the (fast) Scaled Conjugate Gradient [13]. In all cases, the MMSE exponent is smaller than the theoretical value of two, so the power indeed decreases slower, as expected[2].

However, the confined volume of a (usually reverberant) room may also explain the slower decay of the power with the speaker distance. Indeed, if we make an independence assumption between the signal captured from the direct path and the signal captured from an indirect path (reverberation), then the actual received power is the sum of the two powers (direct and indirect).

Section 4.3 provides another verification of the hot air hypothesis, where reverberations have much less interference.

[2]Note that we have neglected the air absorption (exponential decay), because the distances are on the order of 1 or 2 meters only. However, as a sanity check, we estimated the parameters for an (inverse square law + exponential decay) model, but obtained non-realizable parameters (exponential increase instead of exponential decay).



(a) `seq01` (b) `seq02`

Figure 7: The received power decreases slower than the inverse square law. The continuous line is given by the MMSE estimate of the exponent.

4.3 Observation: Elevation Bias

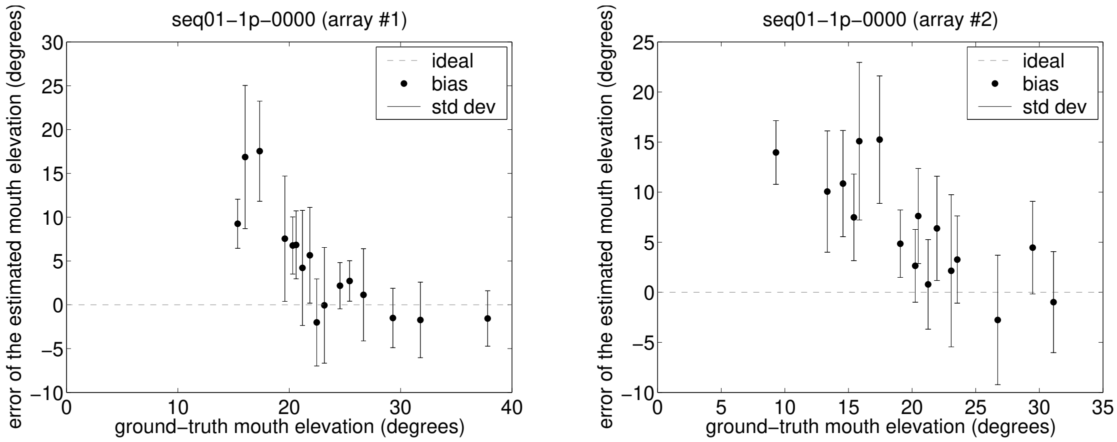
Let us look at the main path, neglecting reverberations. Assuming a "lossy waveguide" of hot air may lead to the main path being curved. This would translate into an important elevation bias of the estimated elevation angle, when the true elevation angle is relatively low but significantly above zero (Fig. 6a). On the contrary, when the true elevation angle is high, the received path is a straight line, and the elevation bias should be close to zero (Fig. 6b). Both points are confirmed by Fig. 8, on both recordings `seq01` and `seq03`, and with both circular arrays (4 experiments). For each recording and for each array, we ran the "FAST" speaker detection-localization algorithm, and the "SNSGMM" speech/non-speech classifier, to extract elevation estimate of the speaker mouth location. Both "FAST" and "SNSGMM" are described in [3].

5 Conclusion

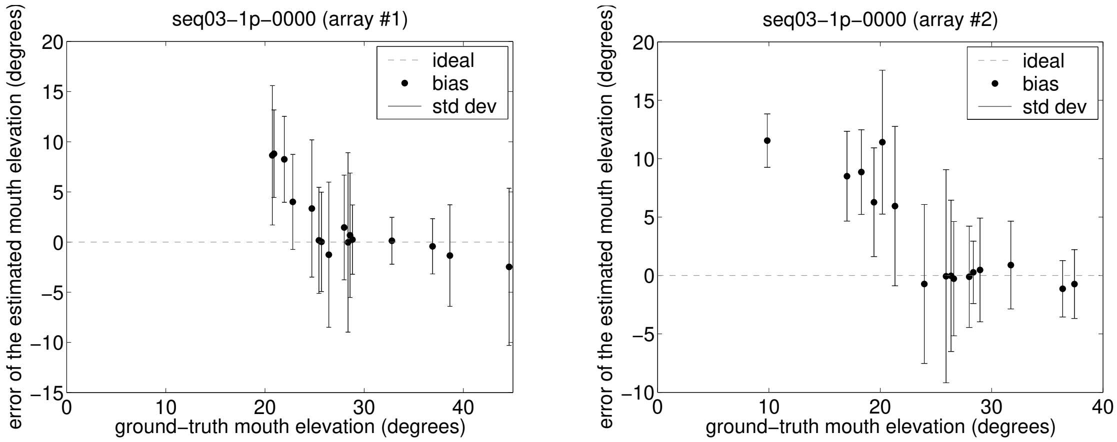
To summarize the investigations conducted so far, it was hypothesized that the hot air stream exhaled by a speaker has a significant impact on the global mouth-to-microphone transmission channel. Observations on two subjects, and 256 pairs (speaker location, microphone location) for each subject, confirm that the hot air stream may indeed have such an impact.

5.1 Multi-Band Asynchrony and Speaker Clustering

Whichever true cause(s) it may have, a firm fact was observed in the above: "multi-band asynchrony". Observations on the long-term average of the complex frequency-domain cross-correlation between a lapel microphone and a distant microphone showed that *different frequencies effectively travel at different speeds*. The speech spectrum can thus be divided in a few bands, where each band has its own Time Of Flight (TOF) from the mouth to the distant microphone. The variation of TOF at the boundary between two such bands appeared to be highly discontinuous (a "jump"). This implies difficulties when using a linear – therefore continuous – process such as [1] to normalize the relative delays between these bands. It is possible that the speaker identity information carried by MFCCs would be much more sensitive to these discontinuities, as compared to the semantic information carried by the same MFCCs.



(a) Recording `seq01`



(b) Recording `seq03`

Figure 8: The error of the estimated elevation depends on the true elevation.

5.2 Future Work

Based on the analyses presented in this paper, further work can focus on at least two directions.

First, further experiments are needed to confirm (or infirm), and precise the *potential cause(s)* of the distance-dependent variability in Time Of Flight, as well as precise its link with the distance-dependent variability in the speaker identity features. A possible experimental protocol could use a variable reverberation chamber, with the same human subject, very still, saying twice the same sentence at each location: once with the room being anechoic, the second time with the room being reverberant. This could permit to identify and possibly separate the contribution in the variabilities, of the reverberations on one hand, and of the hot air stream on the other hand.

Second... Humans can recognize/discriminate speaker identities efficiently at varying distances – certainly up to 2.21 meters! On the other hand, frame-based MFCCs do not seem to have this potential, unless a multi-speaker calibration procedure is used, for each location of a discrete grid covering the room [14]. Humans can obviously adapt to various room configurations without such calibration procedure. For a machine, it is thus desirable to adequately cope with location-dependent non-linearities such as those characterized in terms of TOF. In particular, the observed division of the spectrum in a few bands (for example 1 to 4), each band arriving with a different TOF, suggests multi-band approaches that would be flexible to local asynchrony between the bands. Work in this area includes Asynchronous Hidden Markov Models for multi-modal speech recognition [15], that can handle a local asynchrony between two streams. However, complexity issues arise when handling more than two streams, so further research is needed.

Signal analysis methods that are invariant to location- and frequency-dependent variations of the TOF, may not only benefit to distant speaker clustering, but also to ASR.

Acknowledgments

The author acknowledges the support of the European Union through the AMI and HOARSE projects. This work was also carried out in the framework of the Swiss National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM)2. The author would like to thank Prof. Rainer Martin and Dr Christof Faller for their comments and suggestions.

References

[1] D. Gelbart and N. Morgan, "Evaluating long-term spectral subtraction for reverberant asr," in *Proceedings the IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2001.

[2] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 27, no. 3, pp. 305–317, March 2005.

[3] G. Lathoud, "Further Applications of Sector-Based Detection and Short-Term Clustering," IDIAP-RR-06 26, 2006.

[4] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," IBM T.J. Watson Research Center, Tech. Rep., 1998.

[5] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2003.

[6] G. Lathoud, M. Magimai.-Doss, B. Mesot, and H. Bourlard, "Unsupervised Spectral Subtraction for Noise-Robust ASR," in *Proceedings the IEEE workshop on Automatic Speech Recognition and Understanding (ASRU)*, December 2005.

[7] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, August 1976.

[8] G. Lathoud, J.-M. Odobez, and D. Gatica-Perez, "AV16.3: an Audio-Visual Corpus for Speaker Localization and Tracking," in *Proceedings the workshop on Machine Learning for Multimodal Interfaces (MLMI)*, 2005.

[9] J. Meyer, "Directivity of the bowed stringed instruments and its effect on orchestral sound in concert halls," *Journal of the Acoustical Society of America (JASA)*, vol. 51, no. 6, pp. 1994–2009, 1972.

[10] I. Schwetz, G. Gruhler, and K. Obermayer, "Correlation and stationarity of speech radiation," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, September 2004.

[11] N. J. Shaviv, "Condensation of your exhaled breath," `http://www.sciencebits.com/exhalecondense/`, 2006.

[12] M. Tahira, M. Nomura, Y. Sawada, , and K. Kamo, "Infrasonic and acoustic-gravity waves generated by the Mount Pinatubo eruption of june 15, 1991," in *Fire and Mud: Eruptions and Lahars of Mount Pinatubo, Philippines*, C. Newhall and R. Punongbayan, Eds. Philippine Institute of Volcanology and Seismology, Quezon City and University of Washington Press, Seattle and London, 1996.

[13] M. Moller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, pp. 525–533, 1993.

[14] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speaker recognition based on position dependent cepstral mean normalization," in *Proceedings of Interspeech*, 2005.

[15] S. Bengio, "Multimodal speech processing using asynchronous hidden markov models," *Information Fusion*, vol. 5, no. 2, pp. 81–89, 2004.