Seamless device handover for pervasive speech communication

Vijaya Nirmala Mitnala
University of Essex
Colchester, UK CO4 3SQ
vm20821@essex.ac.uk

Martin J. Reed
University of Essex
Colchester, UK CO4 3SQ
mjreed@essex.ac.uk

Ian Kegel
British Telecommunications plc. (BT)
Adastral Park, Ipswich, UK IP5 3RE
ian.c.kegel@bt.com

John Bicknell
British Telecommunications plc. (BT)
Adastral Park, Ipswich, UK IP5 3RE
john.bicknell@bt.com

***Abstract*—Sustained growth in the smart speaker market has helped establish high quality, far-field speech communications as a viable alternative to the handset. Seamless handover offers a simple but effective way of improving the far-field communication experience by automatically switching to the best available device regardless of where a user is located. While the basic concept of seamless handover has been proven in a lab environment, this paper proposes two significant enhancements: reduction in media disruption during handover by introducing a parallel session on multiple devices through session initiation protocol (SIP) call forking; and, coherence-based signal processing to more accurately determine the most suitable device for the user. The solution proposed uses the magnitude square coherence (MSC) and results verified through simulation and real datasets show it has excellent performance. However, the raw MSC is found to have high variation due to room effects, consequently this work shows that a smoothing predictor is needed to significantly reduce the extraneous transitions that would otherwise be subjectively poor. Unlike a purely location based approach, the proposed solution selects the best smart device without any environment specific calibration making it ideal for straightforward deployment of a pervasive speech application that uses smart speakers.**

Key words– Pervasive communication Speech system, Magnitude Square Coherence, Device handover, Session Initiation Protocol.

I. INTRODUCTION

The Smart Speaker market has seen sustained year-on-year growth since 2015 [**?**] and despite the ongoing semiconductor shortage, over 39 million units were shipped in the third quarter of 2021 alone [**?**]. Whilst much of this growth has been driven by Amazon, Google and Apple, telecommunications providers such as Deutsche Telekom and Orange are actively pursuing strategies to take advantage of smart speaker technology [**?**] and use-cases such as seamless call handover are an obvious area where they can differentiate.

The main objective of seamless handover is to improve the far-field communication experience by removing the need to manually transfer calls between multiple devices. Until quite recently, the only way to switch between devices during a call has been to manually end the session on the active device and start a new one on the target device. Alternatively, a broadcast approach – such as 'group conversation' [**?**] [**?**] - shares any communication session between all devices that are part of the specified group. Whist this makes any device available for the active communication session it means the user has no control over who hears the conversation, raising privacy and usability issues.

As a first step towards improving these experience, a basic seamless call handover experience has been proven in the lab: under the control of a call orchestration script, smart speaker development kits have been used to provide basic user location information and an open-source call server used to perform call session handover between standard Session Initiation Protocol (SIP) clients without any user intervention.

A key requirement in moving from the lab to realistic environments is how well the user tracking works and while volume level and DoA (Direction of Arrival) angle have been sufficient at the Poof of Concept stage, the ability to work in noisy, multi-user environments is a significant challenge.

There are many ways to improve tracking reliability – such as making use of the addition of IoT sensors around the Smart Home – however it is highly desirable to try and derive this information from audio signals available from the devices that make up the communications system itself without requiring input from external devices or precise location calibration. Consequently, this paper addresses this need by using the *coherence* between audio signals in the smart device microphones. Instead of using precise location tracking, our work shows that correct transition between smart devices is possible using the coherence measurement without any form of calibration. Additionally, this work considers how SIP can be used for *mid-call* session handover when using the audio features to identify the suitable device for handover.

The main contributions of this paper are: proposing alternative approaches for SIP mid-call personal mobility; and, using the magnitude square coherence (MSC), together with predictive smoothing, to automatically detect the audio device for session handover.

In Section II we review related literature before describing our proposed methods for both SIP and device detection in III. Our results in IV show the benefit of the proposal, which we comment on in V before we conclude in VI.

II. RELATED WORK

While there are few publications directly addressing smart device switching for pervasive audio, the subject of sound source localisation (SSL) has been studied for various audio signal processing applications such as robotics, video conferencing, smart home systems, speech enhancement, ocean engineering and military systems. SSL involves estimating two parameters of the acoustic signal, namely direction of arrival (DoA) and distance of arrival estimation. Both classical and learning based solutions have been proposed to solve the SSL problem. Classical approachs [**?**] [**?**] require *a priori* knowledge of the acoustic source environment *i.e.,* physical room characteristics such as the room dimension, room impulse response, surface area of the walls *etc*. Hence,

learning based approaches have recently been proposed for localizing using sound distance estimation without these classical measurements [**?**], [**?**], [**?**], [**?**]. The work in [**?**] is a comparative study on calculating distance of arrival based on various acoustic source distance parameters and concludes that the MSC gives the best method for distance calculation. The work in [**?**] describes the use of MSC to calculate the direct-to-reverberant ratio (DRR), a useful parameter in acoustic applications. The work in [**?**] proposes a Gaussian mixture model (GMM) using magnitude square coherence feature from a binaural input for distance of arrival estimation. Their work uses the MSC to estimate distance using *a priori* calibration. While these methods provide good motivation for using the MSC for distance estimation following calibration in static environments, they do not solve the problem of generic device switching where calibration is not available. However, the finding that the MSC is a highly useful metric gives motivation for our work of using the MSC for finding the most suitable device without calibration. Furthermore, these prior works do not consider generic techniques for optimising switching between devices where the position of the devices is not closely controlled, and do not address the signalling required to enable intelligent switching for smart devices.

SIP is widely accepted and heavily used as an application layer signalling protocol in VoIP and IMS systems [**?**]. Advanced Machine Learning (ML) technologies are been researched [**?**], [**?**], [**?**], [**?**], [**?**] to improve the QoS in SIP. The works in [**?**], [**?**], [**?**], [**?**] detail the use of ML with SIP for handling various issues including terminal mobility in vertical handovers and mitigating security issues such as denial of service. The work [**?**] describes approaches for using ML to assist SIP *terminal mobility* for proactive handovers that avoid the handover interruptions in heterogeneous networks. However, while these are important works, they do not address the issue of *personal mobility* support of SIP device handover in homogeneous local area networks, such as those used in home or meeting room environments. Consequently, together with the earlier review of localisation, we have the motivation of this paper to solve both the device detection and signalling problems for these types of networks.

III. Proposed Method

A. SIP signalling for pervasive audio

The standard SIP protocol supports *personal mobility* in addition to *terminal mobility* during *pre-call* or even *mid-call*; although these approaches do not consider seamless media transitions. With *personal mobility*, multiple devices can be registered to one SIP address for session handling. A call session can be initiated on multiple devices with the same SIP address using the existing SIP support of either parallel or sequential call forking. This facility can be used in dynamic session handover between the devices.

In our study we are proposing an out-of-band mechanism, a *pervasive orchestrator*, running on a B2BUA (back-to-back user agent) to detect the suitable device for handover. The device detection is based on the method proposed later in III-B. During a call (i.e. while *mid-call*) on one device, when a more suitable device is detected for handover, the existing session can be transferred to the new device using SIP Re-Invite or REFER methods. The message sequence flow for the case of call sequential forking and using REFER methods is as shown in Fig 1. Note that only the signalling flow is shown in the figure, the message flow related to media is omitted from the diagram. An out-of-band mechanism (the *pervasive orchestrator*) triggers the B2BUA to send a Re-Invite/REFER SIP message to the second device for session handover. The role of the pervasive orchestrator in this case is limited to detecting the device and to make the B2BUA initiate Re-Invite/REFER to the second device; no changes are required to the existing SIP stack. However, previous studies have shown that successful session establishment/re-establishment can take at least 280ms, therefore there will be interruption to the media during session handovers [**?**].

An alternative solution for session handover is proposed in this paper, as shown in Fig 2. This new approach avoids the media interruption and gives better control of the session handling on the devices that are in the handover domain. As shown in Fig 2, the out-of-band *pervasive orchestrator* – running on the B2BUA – will handle the device detection and also the proactive session establishment on all the devices that are within the pervasive application domain. For this session establishment the system will use *parallel call forking*. In the case of *sequential call forking*: when a session is established on one device, the B2BUA does not initiate any SIP messages for session establishment on other devices. However, in the case of *parallel call forking*, the B2BUA, in standard SIP, cancels the Invites that were sent to other devices after one device is connected. However, for this paper we require parallel sessions on all the devices for media continuity, consequently, a change is required in the behaviour of the B2BUA. Specifically, after connecting to one device, the B2BUA should not cancel the Invites sent to other devices. These behavioural changes can be achieved using SIP header extensions or custom headers. It should be noted that the changes only occur inside the handover domain *e.g.,* the smart devices in a home; consequently, remote SIP systems do not need to be modified. The devices connected in the session can be monitored for their health using standard SIP OPTIONS.

B. Nearest device estimation

Here we propose how to perform the intelligent switching between devices within a single audio environment, for example two smart devices (smart speakers) in a room or meeting room. The scenario assumes that a voice call is in progress and that the user wants to automatically swap between smart devices in the audio environment. We will consider only one end of the communication here, although of course the system could be used at both ends of the communication. As described in III-A, the signalling and detection only takes place at one end of the communication so that the other end (or ends in multi-party cases) is unaware.

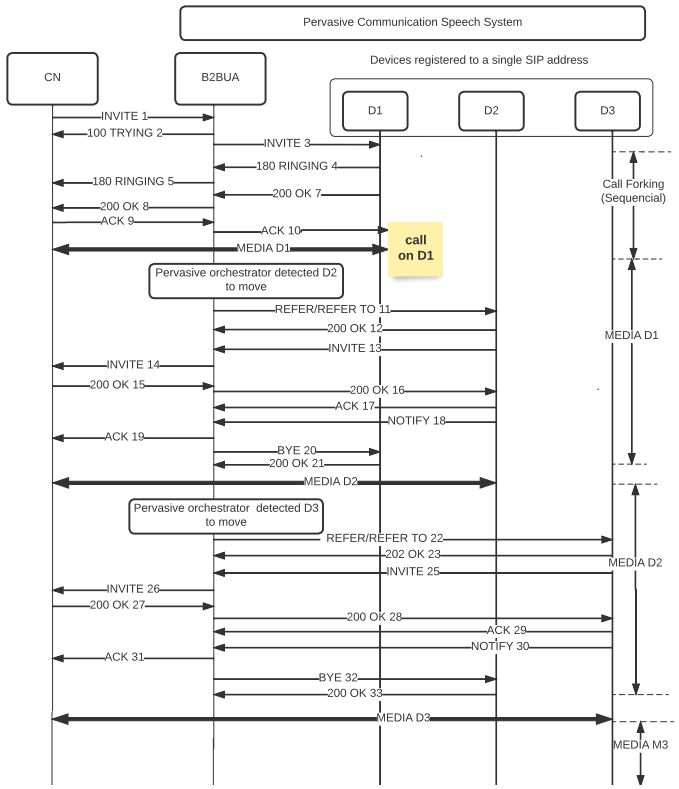

Fig. 1. SIP Session handover-Sequential using forking, note the short breaks in the media transmission.

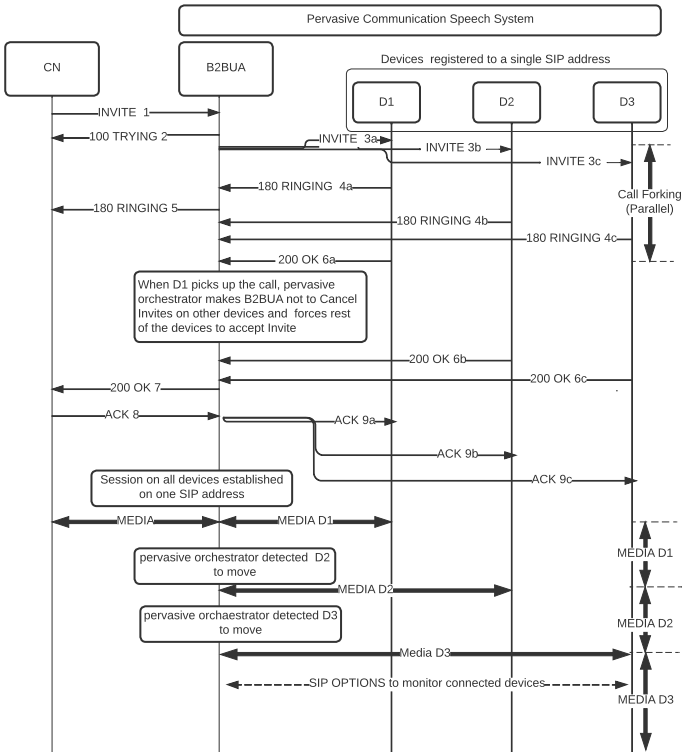

Fig. 2. New Session handover Proposal, note the seamless media transmission during the device transition.

The basic process involves capturing the audio from the smart devices and using the talkers voice to determine which device is best to use for continuing an audio call. While other techniques [**?**] propose using location identification, which usually involves precise calibration and device placement, here we aim to perform the detection using only the talker's voice without knowledge of placement or orientation of the devices. We assume that the devices are equipped with two microphones, as is usual with most smart devices for echo suppression. We also assume that the talker's movement is relatively slow compared to the distances between the smart devices, for example walking from one end of a room to another (*i.e.,* not running between two closely space speakers).

An overview of the proposed procedure is:

- Capture speech separately from the microphone pair from each device (Device 1 and Device 2 in examples later)
- Apply A-weighting to each audio signal to emphasise speech over other background noise
- Calculate the MSC, ρ_1, ρ_2, on a microphone pair from each device over an audio block (see the detailed explanation below)
- Apply a predictor to the output of the MSC over time to remove the effects of noise and low-level transient features due to room acoustics, obtaining the estimates $\hat{\rho_1}$, $\hat{\rho_2}$
- Determine the best device using $\max(\hat{\rho_1}, \hat{\rho_2})$
- Switch the call media to the best device using the signalling described in III-A.

The above steps assume that the process is only taking place while the local user is talking. In practice, voice activity detection is needed to determine when this phase of the conversation is taking place, this is not considered in this paper as this type of detection has been widely considered in the literature [**?**]. We now formally state the procedures.

1) Calculating the Magnitude Square Coherence: The MSC is calculated using Welch's cross-power spectral density for the two microphone signals from device i, $x_{i,l}(t)$ $x_{i,r}(t)$, for the left, l, and right, r, channels respectively. Specifically the MSC $\rho_i(t)$ for a block at time t is calculated using:

$$\rho_i(t) = \frac{\left|\hat{P}(X_l(f,t), X_r(f,t))\right|^2}{\hat{P}(X_l(f,t), X_l(f,t))\hat{P}(X_r(f,t), X_r(f,t))} \quad (1)$$

where the cross power spectral density $\hat{P}(X_a(f,t), X_b(f,t))$ is calculated across an N block Fourier transform of $X_a(f,t)$ from $x(t)$ as:

$$\hat{P}(X_a(f,t), X_b(f,t)) = \frac{1}{N}\sum_{f=0}^{N-1} |X_a^*(f,t)X_b(f,t)| \quad (2)$$

Finally the averaged MSC across M blocks of size K, $t = 0, \ldots K(M-1)$, $\hat{\rho}$ is

$$\hat{\rho} = \frac{1}{M}\sum_{t=0}^{K(M-1)} \rho_i(t) \quad (3)$$

Unfortunately, the MSC calculated using this method is highly dependent on reflections in the room, which cause frequency dependent constructive and destructive interference.

Consequently, we need to predict the MSC $\hat{\rho}$ from a number of previous noisy observations $\rho_t, \rho_{t-1}, \ldots$.

2) Magnitude Square Coherence prediction: The MSC is predicted using the double exponential smoothing method, a popular time series estimator that has been shown to be useful in location tracking applications with considerably lower computation costs than other predictors such as a Kalman filter [**?**]. The algorithm maintains level and trend estimates at each period and prediction of $\hat{\rho}$ is calculated based on weights α and β using:

$$L_t = \alpha\rho_{t-1} + (1-\alpha)(L_{t-1} + T_{t-1}) \quad (4)$$

$$T_t = \beta(L_t - L_{t-1}) + (1-\beta)T_{t-1} \quad (5)$$

$$\hat{\rho}_t = L_t + T_t \quad (6)$$

where L_t represents the *level* and T_t represents the *trend* at time t. The values of α and β will be dependent upon the sampling rate of the MSC, ρ, and typical variations due to movement. In practice it has been found that a Monte-Carlo optimisation with a relatively coarse-grained search allows α and β to be obtained with values that work well.

IV. Results

A. Simulation Setup

A rectangular room with characteristics listed in Table I were simulated using the Pyroomacoustics Python package [**?**]. This room represents a standard living room but without considering artifacts such as variable wall absorbance, windows and furniture. While these additional artifacts would change the specific MSC measured at a single point, this paper is more interested in the relative difference between MSC at the two devices rather than the specific MSC value. In this standard room, the two devices, Device 1 (D1) and Device 2 (D2) are simulated each with two omnidirectional microphones and placed at opposite ends of the room with a distance of 4.5 meters between them. This simulates an environment such as an open-plan living space with say a device in the living room area and another in the kitchen area.

Movement of the sound source (the talker) within the environment is simulated using two methods, as shown in Fig. 3: *linear movement* where the sound source is moved in 0.05 m steps from one end of the room to the other; and, *locus movement* where the sound source is moved in a pseudo-random manner using a set of three random positions (a minimum distance apart) that are connected using a cubic Bezier curve. For the locus movement, a set of 400 loci were created with each locus sampled at 100 equally spaced points along the curve. This simulates the talker moving with a purpose between two locations while navigating around a third object such as furniture. A representative set of six such loci are shown in Fig. 3.

B. Device transition results

The MSC using Welch's method, along with applying A-weighting and double exponential smoothing for prediction, as described in section III-B1, is captured for each position of the sound source on both devices. The device with higher MSC, $\max(\hat{\rho_1}, \hat{\rho_2})$, is considered as the best device to transfer the session. The experiments were performed in simulated rooms with various reverberation time as described by the RT60 metric *i.e.* moderate reverberation case (e.g. living room) to high reverberation conditions (e.g. concert room). The Monte-Carlo parameterisation determined values of $\alpha = 0.05$ and $\beta = 0.01$ as suitable for the double exponential smoothing.

The performance is evaluated using two metrics

number of extraneous transitions: the number of additional device transitions using the predicted talker location compared to the actual nearest device

weighted normalised error (WNE): an error metric $\hat{E}$ normalised to the difference of the distance between the devices and the talker calculated across all the measurements (the set P):

$$\hat{E} = \sum_{i \in P} \frac{e_i \, |d_i|}{D} \quad (7)$$

where $D = \sum_{i \in P} |d_i|$; $|d_i|$ is the absolute difference between the talker and Device 1 and the talker and Device 2; and e_i is a 0,1 variable which is zero if the predicted device is the same as the nearest device, or one otherwise.

The use of extraneous transitions as a metric is important as users would find switching that is unnecessary as subjectively disturbing, the ideal is for this to be zero. The purpose of the WNE, $\hat{E}$, is to compute an error that is higher if the device chosen is significantly further away than the ideal device.

The results of the linear movement scenario compared to the ideal transition, and for various simulated room RT60 values, is shown in Table II. Additionally, examples of the output of the MSC output are shown in Figures 4 and 5 for RT60 of 0.6 and 1.5 respectively. The MSC values (coh. on the graph) in Figures 4 and 5 show how the raw MSC values can vary significantly, with artifacts present due to additive reflections at certain positions; these are clearly worse with the higher RT60 value used for Figure 5. The NS (non-smoothed) data in the figures shows how the transitions cannot clearly be detected using the raw MSC, whereas the transition in the S (smoothed) are very close to the ideal. It can be seen that the transition happens at not quite the optimal location if purely distance is used, however, in practice this is not likely to be subjectively important as users would want the device to switch appropriately rather than expecting millimeter precision on their location being exactly equidistant from the devices. Again the NS curve shows that, without the predictive smoothing, there would be many extraneous transitions that would be subjectively annoying. The results in Table II show that the proposed technique (MSC plus predictive smoothing) work very well compared to raw MSC results, which have an increasingly large number of extraneous transitions and error (WNE) as the reverberation increases.

We also tested our work on more complex *locus* movement as described in IV-A and examples shown in Fig. 3. The results of the extraneous transitions for these scenarios is shown in Fig. 6 showing that it was a more challenging problem;

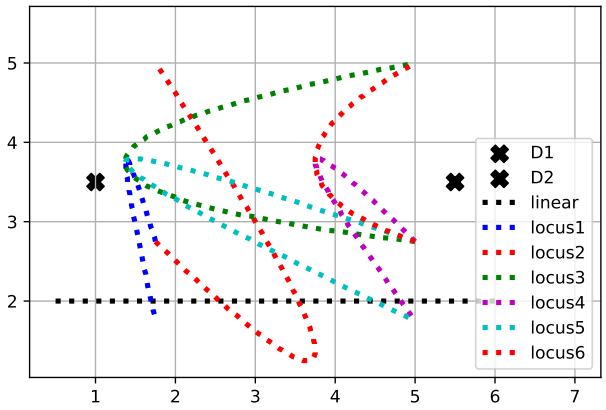

Fig. 3. Linear and locus source positions compared to fixed device positions (D1, D2). A sample of six loci from the 400 generated are shown.

TABLE I
SIMULATION SETUP

Parameter	Description
Room Dimensions	7m X 5.5m X 2.4m
Device positions	Device1: 1.0m X 3.5m X 0.9m Device2: 5.5m X 3.5m X 0.9m
Source Positions	110 positions starting from 0.5m to 6m each with 0.05m steps
RT60	0.6s, 0.8s, 1.0s, 1.5s, 1.8s
Audio sample rate	16 kHz
Audio block size	1024 samples
Audio block averaging	16,384 (16 blocks)

however, we can see that the number of extraneous transitions is considerably lower when using the MSC with predictive smoothing (S on the graph) compared to the raw MSC values (NS on the graph).

Our solution was also verified on a real room/talker dataset obtained from [?] [?] where it achieved 98% accuracy in identifying the suitable device for a given talker location.

V. DISCUSSION

The results show that the method proposed can work very well for detecting the appropriate device to switch a

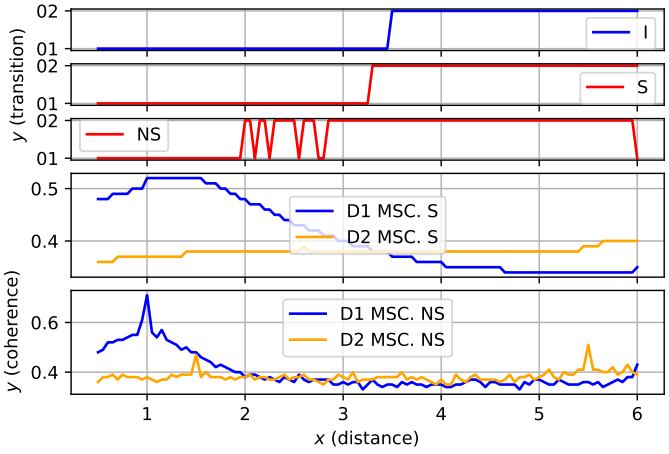

Fig. 4. Device transition for Reverberation Time $RT60 = 0.6s$ showing ideal device transition (I), our, smoothed, solution (S), the non-smoothed case (NS) and the relative MSC values used to derive transitions.

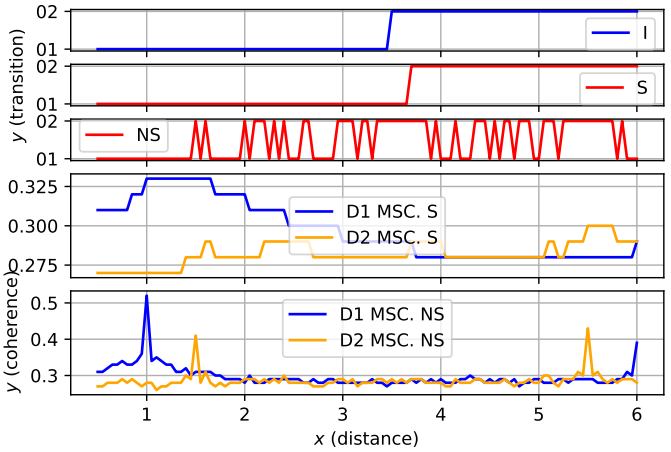

Fig. 5. Device transition for Reverberation Time $RT60 = 1.5s$ showing ideal device transition (I), our, smoothed, solution (S), the non-smoothed case (NS) and the relative MSC values used to derive transitions.

TABLE II
COMPARISON BETWEEN DIFFERENT SIMULATION SCENARIOS USING PROPOSED METRICS OF WEIGHTED-NORMALISED ERROR (WNE) AND NUMBER OF EXTRANEOUS TRANSITIONS (ET)

	RT60				
Type	0.6s	0.8s	1.0s	1.5s	1.8s
WNE-NS	0.14	0.20	0.24	0.25	0.3
WNE-S	0	0.01	0.01	0.01	0.18
ET-NS	9	17	22	38	38
ET-S	0	0	0	0	0

session to when two smart devices are part of a pervasive speech session. The method proposed does not use any actual location information or calibration of the device orientation. The metrics in the evaluation used a measure of error (the WME $\hat{E}$) that determined a transition as an error if it did not occur at the equidistant point between the speakers, this was weighted by the "error" in the distance. However, in practice it is unlikely that users will be concerned about the actual switching point being at the equidistant point, an error of a few tens of centimeters or even a meter or two may be perfectly adequate. Indeed consider a situation of an open-

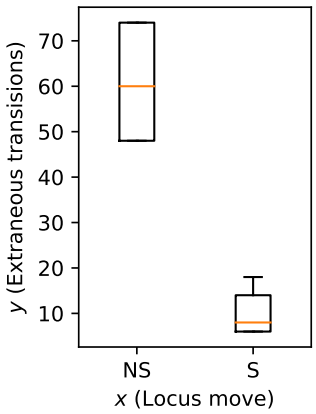

Fig. 6. Extraneous transitions in Smooth (S) and in No smooth (NS) case for locus move

plan living environment with one smart device placed in a relatively acoustically dead living area and another placed in an acoustically live kitchen area. In this scenario it may be beneficial for the switching to occur closer to the kitchen area so that the time spent using the device in the kitchen area is minimised to the user being located close to the smart device in this area. If the switching was done purely on distance (*i.e.,* actual location) then the non-ideal kitchen device might be used when it would be preferable to use the device in the living area. The use of the MSC, as proposed in this paper, conveniently obviates this problem as the MSC will tend to be less in acoustically live areas and higher in acoustically dead areas, such as the living room space of the example. Consequently, it is proposed that the method presented in this paper selects a "preferable" device rather than simply the "nearest" device; however, further subjective testing will be required to confirm this hypothesis.

VI. CONCLUSIONS

This paper proposes seamless *mid-call* session handover between the audio devices in a pervasive communication application using the SIP protocol and coherence based audio signal processing for detecting the correct smart device. The analysis shown that MSC, along with a smoothing predictor, can provide very accurate prediction of the optimum smart device for communication. The results show that a smoothing predictor greatly reduces extraneous transitions and transitions with greater accuracy compared to using the raw MSC values. The proposed method was verified in various room conditions for two simulated devices and also for a number of devices using a real room/talker dataset.