



A multi-microphone approach to speech processing in a smart-room environment

Abad Garetà, Alberto

Use this identifier to quote or link this thesis: <http://hdl.handle.net/10803/6906>

Director: Hernando Pericás, Francisco Javier

Department/Institute: Universitat Politècnica de Catalunya. Departament de Teoria del Senyal i Comunicacions

Date of defense: 2007-06-29

Legal Deposit: B.51286-2007

ISBN: 9788469089705

Keywords: localització de fonts acústiques; detecció d'orientació; processament multi-micròfon; arrays de micròfons; reconeixement de parla; processament de la parla; sales intel·ligents

Subject(s): 621.3 - Enginyeria elèctrica. Electrotècnia. Telecomunicacions

Abstract:

Els avenços recents en tecnologia informàtica i processament de la parla i del llenguatge, entre altres, han fet possible que noves maneres de comunicació entre les persones i les màquines comencin a semblar factibles. Concretament, l'interès en el desenvolupament de noves aplicacions en entorns tancats equipats amb múltiples sensors multimodals, també coneguts com sales intel·ligents, ha augmentat considerablement darrerament.

En general, és ben conegut que la qualitat de les senyals de la parla capturades per micròfons que poden trobar-se a diversos metres de distància dels locutors es veu severament degradada pel soroll acústic i per la reverberació de la sala. En el context del desenvolupament d'aplicacions de la parla en entorns de sales intel·ligents, l'ús de sensors que no siguin molestos és un requeriment habitual. És a dir, no està normalment permès o no és possible fer servir micròfons propers o de solapa, i per tant, les tecnologies de la parla desenvolupades han de basar-se en les senyals capturades per micròfons llunyans. En aquestes situacions, les tecnologies de la parla que habitualment funcionen raonablement bé en entorns lliures de soroll i reverberació pateixen una davallada dràstica en les seves prestacions. En aquesta tesi s'investiguen mètodes multi micròfon per a solucionar els problemes que provoca l'ús de micròfons llunyans en les aplicacions de la parla que habitualment es desenvolupen en sales intel·ligents. Concretament, s'estudia el processament d'arrays de micròfons com a un mètode possible d'aprofitar la disponibilitat de múltiples micròfons per a obtenir senyals de veu millorades. Mitjançant la correcta combinació de les senyals que incideixen en una agrupació de micròfons, el processament d'arrays permet apuntar direccions espacials concretes a l'hora que altres es rebutgen.

Per a la millora de la parla amb arrays de micròfons, en la tesi es proposa l'ús d'un nou esquema robust de conformació que integra en només etapa un conformador adaptatiu i una etapa de post-filtrat de Wiener. Els resultats obtinguts mostren que el conformador proposat és una solució adequada per a entorns molt sorollosos i que, en general, és preferible a l'ús convencional d'etapes de post-filtrat a la sortida d'un conformador adaptatiu. No obstant això, el conformador mostra una certa degradació de la senyal de veu que pot afectar a la seva utilitat per a aplicacions de reconeixement de la parla, especialment quan el soroll no és massa important.

A continuació s'investiga l'ús específic d'arrays de micròfons per al reconeixement de la parla en entorns de sales intel·ligents. Es demostra que l'ús convencional d'arrays de micròfons per al reconeixement de la parla, que consisteix en la seva aplicació en dues etapes independents, no aporta una millora significativa respecte de l'ús de tècniques mono-canal, especialment, si el reconeixedor està adaptat a les condicions reals de l'entorn acústic. En la tesi es fa èmfasi en la necessitat de que el reconeixement de la parla incorpori informació de la conformació amb arrays de micròfons, o alternativament, que els conformadors incorporin informació del reconeixement de la parla. Més concretament, es proposa utilitzar les dades primer capturades per un array de micròfons i després processades per un conformador per a la construcció dels models acústics, per a d'aquesta manera, obtenir un major benefici dels arrays de micròfons. La aplicació del esquema proposat d'adaptació amb dades conformades d'un array, permet obtenir una millora considerable en un sistema de reconeixement depenent de locutor, mentre que en el cas d'un sistema independent de locutor només s'obté una millora molt limitada, degut en part a l'ús de dades d'array simulades.

Per altra banda, una limitació habitual del rocessament d'arrays de micròfons és que es necessita una estimació versemblant de la posició del locutor per a poder apuntar correctament cap a la posició d'interès. A més, el coneixement de la posició de les fonts acústiques que poden estar presents en una sala és una informació que pot ser aprofitada per altres serveis que es desenvolupen en les sales intel·ligents, com per exemple per a apuntar automàticament una càmera en vídeo-conferències. Afortunadament, existeixen nombrosos mètodes que permeten solucionar el problema del seguiment de fonts acústiques basant-se en les senyals capturades per múltiples micròfons.

Concretament, a la tesi es desenvolupa un sistema robust de localització de locutor basat en un dels algorismes

Search

☒ All of TDX

☐ This Department

Consult

All of TDX

- By universities & departments
- By defense date
- By authors/directors
- By titles
- By subjects

This Department

- By defense date
- By authors/directors
- By titles
- By subjects

Statistics

By these

Other Theses Websites

European Theses (DART)
International Theses



Finalment, a la tesi també s'estudia el problema de l'estimació de l'orientació del locutor en base a les senyals rebudes per múltiples micròfons, en el context del desenvolupament de noves tecnologies que poden aportar informació addicional per als sistemes que potencialment poden actuar en sales intel·ligents. En concret, es proposen i comparen dos mètodes completament diferents. Per una banda, mètodes sofisticats basats en l'estimació conjunta de la posició i de l'orientació permeten assolir estimacions acceptables a canvi d'un elevat cost computacional. Per altra banda, els mètodes més simples que es basen en consideracions sobre el diagrama de radiació de la parla encara que no són capaços d'assolir les prestacions dels mètodes sofisticats, també poden resultar adequats en alguns casos, como ara quan es coneix la posició amb antelació, o bé quan la despesa computacional està limitada. En tots dos casos, els resultats obtinguts permeten ser optimistes de cara al futur desenvolupament de nous algorismes adreçats a l'estimació de l'orientació del locutor.

Los avances recientes en tecnología informática y procesamiento del habla y del lenguaje, entre otros, han hecho posible que nuevos modos de comunicación entre las personas y las máquinas empiecen a parecer factibles. Concretamente, el interés en el desarrollo de nuevas aplicaciones en entornos cerrados equipados con múltiples sensores multimodales, también conocidos como salas inteligentes, ha aumentado considerablemente en los últimos tiempos.

En general, es bien sabido que la calidad de las señales de habla capturadas por micrófonos que pueden encontrarse a varios metros de distancia de los locutores se ve severamente degradada por el ruido acústico y por la reverberación de la sala. En el contexto del desarrollo de aplicaciones del habla en entornos de salas inteligentes, el uso de sensores que no sean molestos es un requisito habitual. Es decir, normalmente no está permitido o no es posible usar micrófonos cercanos o de solapa, y por lo tanto, las tecnologías del habla desarrolladas tienen que basarse en las señales capturadas por micrófonos lejanos. En estas situaciones, las tecnologías del habla que habitualmente funcionan razonablemente bien en entornos libres de ruido y reverberación sufren un descenso drástico en sus prestaciones.

En esta tesis se investigan métodos multi micrófono para solventar los problemas que provoca el uso de micrófonos lejanos en las aplicaciones del habla que habitualmente se desarrollan en salas inteligentes. Concretamente, se estudia el procesamiento de arrays de micrófonos como un método posible de aprovechar la disponibilidad de múltiples micrófonos para obtener señales de voz mejoradas. Mediante la correcta combinación de las señales que inciden en una agrupación de micrófonos, el procesamiento de arrays permite apuntar direcciones espaciales concretas a la vez que otras se rechazan.

Para la mejora del habla con arrays de micrófonos, en la tesis se propone el uso de un nuevo esquema robusto de conformación que integra en una sola etapa un conformador adaptativo y una etapa de post-filtrado de Wiener. Los resultados obtenidos muestran que el conformador propuesto es una solución adecuada para entornos muy ruidosos y que, en general, es preferible al uso convencional de etapas de post-filtrado a la salida de un conformador adaptativo. Sin embargo, el conformador muestra cierta degradación de la señal de voz que puede afectar a su utilidad para aplicaciones de reconocimiento del habla, especialmente cuando el ruido no es demasiado importante.

A continuación se investiga el uso específico de arrays de micrófonos para el reconocimiento del habla en entornos de salas inteligentes. Se demuestra que el uso convencional de arrays de micrófonos para reconocimiento del habla, que consiste en su aplicación en dos etapas independientes, no aporta una mejora significativa respecto al uso de técnicas mono canal, especialmente, si el reconocedor está adaptado a las condiciones reales del entorno acústico. En la tesis se hace énfasis en la necesidad de que el reconocimiento del habla incorpore información de la conformación con arrays de micrófonos, o alternativamente, que los conformadores incorporen información del reconocimiento del habla. Más concretamente, se propone el uso de datos capturados por un array de micrófonos y luego procesados por un conformador para la construcción de los modelos acústicos, para de esta manera, obtener un mayor beneficio de los arrays. La aplicación del esquema propuesto de adaptación con datos conformados de un array de micrófonos permite obtener una mejora considerable en un sistema de reconocimiento dependiente de locutor, mientras que en el caso de un sistema independiente de locutor sólo se obtiene una mejora muy limitada, debido en parte al uso de datos de array simulados.

Por otro lado, una limitación habitual del procesamiento de arrays de micrófonos es que se necesita una estimación verosímil de la posición del locutor para poder apuntar correctamente hacia la posición de interés. Además, el conocimiento de la posición de las fuentes acústicas que puedan estar presentes en una sala es una información que puede ser aprovechada por otros servicios que se desarrollan en las salas inteligentes, como por ejemplo para apuntar automáticamente una cámara en video-conferencias. Afortunadamente, existen numerosos métodos que permiten resolver el problema del seguimiento de fuentes acústicas basándose en las señales capturadas por múltiples micrófonos.

Concretamente, en la tesis se desarrolla un sistema robusto de localización de locutor basado en uno de los algoritmos actuales de mayor éxito consistente en el cómputo de la verosimilitud de cada posible posición basándose en las estimaciones de las correlaciones cruzadas generalizadas entre pares de micrófonos. El sistema propuesto incorpora principalmente dos novedades. Primero, las correlaciones cruzadas se calculan de forma adaptativa basándose en las velocidades estimadas de las fuentes. Este cálculo adaptativo se hace de manera que se minimice el efecto de las diferentes dinámicas de las fuentes presentes en la sala en el resultado de la localización. Segundo, se propone el uso de un método acelerado para el cálculo de la posición basado en estrategias de búsqueda de menor a mayor resolución tanto en el dominio espacial como frecuencial. De hecho, se muestra que la relación entre resolución espacial y el ancho de banda considerado en el cálculo de las correlaciones cruzadas es un aspecto fundamental a tener en cuenta en la aplicación adecuada de este tipo de estrategias rápidas. Las dos novedades comentadas permiten que el sistema propuesto alcance unos resultados razonablemente buenos cuando se evalúa en escenarios relativamente controlados y con pocos locutores que no se solapan. Además, la conveniencia del sistema de localización acústica propuesto queda de manifiesto si se atiende a los destacados resultados que se obtuvieron en una evaluación internacional.

Finalmente, en la tesis también se estudia el problema de la estimación de la orientación del locutor en base a las señales capturadas por múltiples micrófonos en el contexto del desarrollo de nuevas tecnologías que puedan aportar información adicional para los sistemas que potencialmente pueden actuar en salas inteligentes. En concreto, se proponen y comparan dos métodos completamente diferentes. Por un lado, métodos sofisticados basados en la estimación conjunta de la posición y de la orientación que permiten obtener estimaciones aceptables a cambio de un elevado coste computacional. Por otro lado, los métodos más simples que se basan en consideraciones sobre el diagrama de radiación del habla aunque no son capaces de igualar las prestaciones de los métodos sofisticados,

talking microphones is usually not allowed, and consequently, speech technologies must operate on the basis of distant-talking recordings. In such conditions, speech technologies that usually perform reasonably well in free of noise and reverberation environments show a dramatically drop of performance.

In this thesis, the use of a multi-microphone approach to solve the problems introduced by far-field microphones in speech applications deployed in smart-rooms is investigated. Concretely, microphone array processing is investigated as a possible way to take advantage of the multi-microphone availability in order to obtain enhanced speech signals. Microphone array beamforming permits targeting concrete desired spatial directions while others are rejected, by means of the appropriate combination of the signals impinging a microphone array.

A new robust beamforming scheme that integrates an adaptive beamformer and a Wiener post-filter in a single stage is proposed for speech enhancement. Experimental results show that the proposed beamformer is an appropriate solution for high noise environments and that it is preferable to conventional post-filtering of the output of an adaptive beamformer. However, the beamformer introduces some distortion to the speech signal that can affect its usefulness for speech recognition applications, particularly in low noise conditions.

Then, the use of microphone arrays for specific speech recognition purposes in smart-room environments is investigated. It is shown that conventional microphone array based speech recognition, consisting on two independent stages, does not provide a significant improvement with respect to single microphone approaches, especially if the recognizer is adapted to the actual acoustic environmental conditions. In the thesis, it is pointed out that speech recognition needs to incorporate information about microphone array beamformers, or otherwise, beamformers need to incorporate speech recognition information. Concretely, it is proposed to use microphone array beamformed data for acoustic model construction in order to take more benefit from microphone arrays. The result obtained with the proposed adaptation scheme with beamformed enrollment data shows a remarkable improvement in a speaker dependent recognition system, while only a limited enhancement is achieved in a speaker independent recognition system, partially due to i

ii the use of simulated microphone array data.

On the other hand, a common limitation of microphone array processing is that a reliable speaker position estimation is needed to correctly steer the beamformer towards the position of interest. Additionally, knowledge about the location of the audio sources present in a room is information that can be exploited by other smart-room services, such as automatic video steering in conference applications. Fortunately, audio source tracking can be solved on the basis of multiple microphone captures by means of several different approaches.

In the thesis, a robust speaker tracking system is developed based on successful state of the art SRP-PHAT algorithm, which computes the likelihood of each potential source position on the basis of the generalized cross-correlation estimations between pairs of microphones. The proposed system mainly incorporates two novelties: firstly, cross-correlations are adaptively computed based on the estimated velocities of the sources. The adaptive computation permits minimizing the influence of the varying dynamics of the speakers present in a room on the overall localization performance. Secondly, an accelerated method for the computation of the source position based on coarse-to-fine search strategies in both spatial and frequency dimensionalities is proposed. It is shown that the relation between spatial resolution and cross-correlation bandwidth is a matter of major importance in this kind of fast search strategies. Experimental assessment shows that the two novelties introduced permit achieving a reasonably good tracking performance in relatively controlled environments with few non-overlapping speakers. Additionally, the remarkable results obtained by the proposed audio tracker in an international evaluation confirm the convenience of the algorithm developed.

Finally, in the context of the development of novel technologies that can provide additional cues of information to the potential services deployed in smart-room environments, acoustic head orientation estimation based on multiple microphones is also investigated in the thesis. Two completely different approaches are proposed and compared: on the one hand, sophisticated methods based on the joint estimation of speaker position and orientation are shown to provide a superior performance in exchange of large computational requirements. On the other hand, simple and computationally cheap approaches based on speech radiation considerations are suitable in some cases, such as when computational complexity is limited or when the source position is known beforehand. In both cases, the results obtained are encouraging for future research on the development of new algorithms addressed to the head orientation estimation problem.

Rights: ADVERTIMENT. L'accés als continguts d'aquesta tesi doctoral i la seva utilització ha de respectar els drets de la persona autora. Pot ser utilitzada per a consulta o estudi personal, així com en activitats o materials d'investigació i docència en els termes establerts a l'art. 32 del Text Refós de la Llei de Propietat Intel·lectual (RDL 1/1996). Per altres utilitzacions es requereix l'autorització prèvia i expressa de la persona autora. En qualsevol cas, en la utilització dels seus continguts caldrà indicar de forma clara el nom i cognoms de la persona autora i el títol de la tesi doctoral. No s'autoritza la seva reproducció o altres formes d'explotació efectuades amb finalitats de lucre ni la seva comunicació pública des d'un lloc allè al servei TDX. Tampoc s'autoritza la presentació del seu contingut en una finestra o marc allè a TDX (framing). Aquesta reserva de drets afecta tant als continguts de la tesi com als seus resums i índexs.

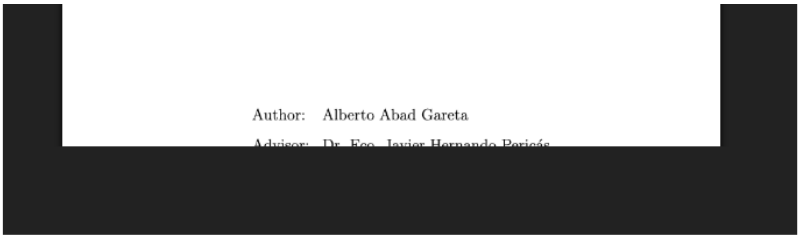
Share:

Download:

DIDL MARC MARC_CCUC METS OAI_DC ORE QDC RDF

Show full item record





Author: Alberto Abad Garetà
Advisor: Dr. Ego Javier Hernandez Beriofé

Files in this item

01Aag01de01.pdf (3.757Mb) 

This item appears in the following Collections

- Departament de Teoria del Senyal i Comunicacions [309]