Resum

Aquesta tesi s'emmarca en el camp del processament de senyals acústics i les seves aplicacions per a entorns de comunicació emergents. El processat de senyals acústics és una àrea de recerca molt àmplia que abasta el disseny d'algorismes per al tractament d'un o diversos senyals acústics per tal de realitzar una tasca determinada, com pot ser: la localització de la font de so que va originar els senyals acústics aconseguits, la millora de la relació senyal a soroll de les mateixes, la separació de senyals d'interès a partir d'un conjunt de fonts interferents o el reconeixement del tipus de font i / o el contingut del missatge. Entre les tasques anteriors, la localització de font de sons (SSL, *Sound Source Localization*) i el reconeixement automàtic de veu (ASR, *Automatic Speech Recognition*) han estat especialment tractades en aquesta tesi. De fet, la localització de fonts de so en una habitació ha rebut molta atenció per part de la comunitat científica en les últimes dècades. La majoria de les aplicacions reals d'arrays de micròfons necessiten localitzar una o més fonts de so actives en condicions adverses (baixa relació senyal-soroll i una alta reverberació). Algunes d'aquestes aplicacions són els sistemes de teleconferència, videojocs, robots autònoms, sistemes remots de vigilància, l'adquisició de senyal en mode mans lliures, etc. De fet, la localització robusta de fonts de so sota condicions d'alt nivell de soroll i reverberació segueix sent un repte. Un dels algorismes més coneguts per a la localització de fonts en entorns sorollosos i reverberants és el *Steered Response Power - Phase Transform* (SRP-PHAT), que constitueix el marc de referència per a les contribucions que es proposen en aquesta tesi. Un altre desafiament en el disseny d'algorismes de SSL és aconseguir el seu funcionament en temps real amb una alta precisió en la localització i amb un nombre raonable de micròfons a un cost computacional reduït. Encara que el algoritme SRP-PHAT ha demostrat ser un algorisme de localització efectiu en entorns reals, la seva aplicació pràctica es basa en general en un procediment costós de recerca per mallat, pel que el cost computacional d'aquest mètode suposa un problema a considerar. És per això que diverses modificacions i optimitzacions s'han proposat en la literatura per millorar el seu rendiment i aplicabilitat. En aquesta tesi es proposa una nova estratègia que estén eficaçment el comportament de l'algorisme SRP-PHAT convencional. Aquest nou mètode realitza una exploració completa de l'espai mostrejat en lloc de calcular el SRP en posicions espacials discretes, augmentant així la seva robustesa i permetent un mallat espacial més ample que el cost computacional requerit en una aplicació pràctica, reduint també el cost en hardware (menor nombre de micròfons). Aquesta estratègia permet implementar aplicacions en temps real basant-se en la informació de les posicions estimades, com ara redirigir de forma automàtica la posició d'una càmera o la detecció de fragments de parla / no parla per a sistemes avançats de videoconferència.

Com s'ha comentat anteriorment, a més de les contribucions relacionades amb SSL, aquesta tesi està també relacionada amb el camp del reconeixement automàtic de veu (ASR). Aquesta tecnologia permet a un ordinador o dispositiu electrònic identificar les paraules pronunciades per una persona perquè el missatge es pugui emmagatzemar i processar d'una forma útil. ASR és utilitzat en el dia a dia en una sèrie d'aplicacions i serveis, com a interfaces home-màquina naturals, sistemes de dictat, traductors electrònics i taulells d'informació automàtica. No obstant això, encara hi ha alguns reptes que cal resoldre. Un problema important en ASR és reconèixer a les persones que estan parlant en una habitació mitjançant l'ús de micròfons a distància. En el reconeixement de veu distant, els micròfons no només reben el senyal via directa de les fonts de so, sinó que també reben rèpliques retardades com a resultat de la propagació multitrajecte. D'altra banda, també hi ha múltiples situacions en les teleconferències en què diversos oradors parlen simultàniament. En aquest context, quan múltiples senyals de veu són presents simultàniament, els mètodes de separació de fonts de so (SSS, *Sound Source Separation*) es poden utilitzar amb

èxit per millorar el rendiment del reconeixement automàtic de veu en escenaris amb múltiples fonts. Amb l'objectiu de millorar aquest tipus de situacions, en aquesta tesi s'ha proposat un mètode d'entrenament diferent. Aquest entrenament, el qual es basa en un model robust construït a partir de veus prèviament separades en diversos entorns acústics, utilitza les tècniques de separació com una etapa de millora de la parla que suprimeix les interferències no desitjades. S'ha estudiat la combinació de la separació de fonts i l'ús d'aquest entrenament específic per a la millora del reconeixement de veu en diferents condicions acústiques, donant lloc a millores de fins a un 35% en la taxa final de reconeixement.

Paraules Clau: Localització de fonts de so, separació de fonts de so, SRP-PHAT, array de micròfons, detecció de parla, reconeixement automàtic de veu.