نشر نویسه پارسی

مجموعه مقالات سومین همایش ملی

زبان شناسی پیکره ای

به کوشش: آزاده میرزائی

مجموعه مقالات سومین همایش ملی

زبان‌شناسی پیکره‌ای

به کوشش

دکتر آزاده میرزائی

(عضو هیئت علمی دانشگاه علامه طباطبائی)

سرشناسه	همایش ملی زبان‌شناسی پیکره‌ای (سومین : ۱۳۹۸ : تهران)
عنوان و نام پدیدآور	مجموعه مقالات سومین همایش ملی زبانشناسی پیکره‌ای/به کوشش آزاده میرزایی ؛ برگزارکننده انجمن زبانشناسی ایران ، با همکاری پژوهشگاه علوم انسانی و مطالعات فرهنگی ، پایگاه استنادی علوم جهان اسلام (ISC) ، نشر نویسه پارسی.
مشخصات نشر	تهران : نشر نویسه پارسی، ۱۳۹۸.
مشخصات ظاهری	۳۳۷ ص.: مصور، جدول، نمودار.
شابک	۹۷۸-۶۲۲-۶۶۴۹-۲۲-۳
وضعیت فهرست نویسی	فیپا
موضوع	زبان‌شناسی پیکره‌ای -- کنگره‌ها
موضوع	Corpus (Linguistics) -- Congresses
موضوع	زبان‌شناسی -- ایران -- کنگره‌ها
موضوع	Linguistics -- Iran -- Congresses
شناسه افزوده	میرزائی، آزاده، ۱۳۵۸ - ، گردآورنده.
شناسه افزوده	انجمن زبان‌شناسی ایران.
رده بندی کنگره	۱۳۹۸ ۳۸۸ز / P۱۲۸
رده بندی دیویی	۴۱۰/۱۸۸
شماره کتابشناسی ملی	۵۶۳۶۴۵۱

انجمن زبان شناسی ایران
تهران، بزرگراه چمران، پل مدیریت، خیابان علامه طباطبایی جنوبی، دانشکده ادبیات فارسی و زبان‌های خارجی دانشگاه علامه طباطبایی، طبقه اول، اتاق ۱۱۷. صندوق پستی: ۱۵۹۷۶۳۳۱۱۱
تلفن:۸۸۶۹۰۰۲۲ نمابر: ۸۸۶۹۰۰۲۲
www.lsi.ir

نشر نویسه پارسی
Neveeseh Parsi Publishing
تهران، صندوق پستی ۱۳۷۹-۱۶۷۶۵
تلفن:۷۷۰۵۳۲۴۶
فروشگاه:۶۶۴۶۱۰۰۷
سامانه پیام کوتاه:۳۰۰۰۴۵۵۴۵۵۴۱۴۲
وبگاه نشر نویسه پارسی:
www.neveeseh.com

شابک: ۳- ۲۲- ۶۶۴۹- ۶۲۲-۹۷۸
ISBN: 978-622-6649-22-3

مجموعه مقالات سومین همایش ملی زبان‌شناسی پیکره‌ای

به کوشش	آزاده میرزایی
ویراستار چکیده‌های انگلیسی	معصومه استاجی
	(عضو هیئت علمی دانشگاه علامه طباطبائی)
طرح جلد و یونیفورم	محمد محرابی www.mehraabi.com
صفحه‌آرایی و آماده‌سازی چاپ	محمد محرابی www.mehraabi.com
چاپ و صحافی	روز
شمارگان	۳۰۰ نسخه
نوبت چاپ	اول، ۱۳۹۸
قیمت	۵۵۰۰۰ تومان

مجموعه مقالات سومین همایش ملی زبان‌شناسی پیکره‌ای

برگزارکننده

انجمن زبان‌شناسی ایران

با همکاری

پژوهشگاه علوم انسانی و مطالعات فرهنگی

پایگاه استنادی علوم جهان اسلام (ISC)

نشر نویسه پارسی

هیئت داوران

کمیته اجرایی

فهرست مطالب

پیشگفتار

سومین همایش ملی زبان‌شناسی پیکره‌ای، در اردیبهشت ۱۳۹۸ در پژوهشگاه علوم انسانی و مطالعات فرهنگی به همت انجمن زبان‌شناسی ایران برگزار شد. برای این همایش ۲۰ مقاله از دانشگاه‌ها و مراکز تحقیقاتی سراسر ایران به دبیرخانهٔ همایش ارسال شد که از میان آنها، پس از داوری، ۱۵ مقاله برای سخنرانی و چاپ برگزیده شد.

کارآمدی پیکره‌ها در پژوهش‌های نظری و کاربردی مرتبط با زبان، سبب شده است تا ساخت پیکره‌ها از یک سو و روش‌ها و رویکردها در بهره‌برداری از پیکره‌ها با اهداف پژوهشی متنوع، در سوی دیگر موردتوجه پژوهشگران حوزه‌های مختلف قرار بگیرد. امروزه زبان‌شناسی پیکره‌ای چه در مقام ابزاری پژوهشی که بررسی پدیده‌ها و نظریات مطرح در زبان را هدف قرار می‌دهد و چه در جایگاهی که صدور فرضیات و نظریات زبانی را رقم می‌زند، جزئی جدایی‌ناپذیر از پژوهش‌های زبانی است. در ارتباط با زبان فارسی در حال حاضر دستاوردهای ارزنده‌ای هم به لحاظ کمی و سرعت پیشرفت محتوا و هم از جهت تعداد، در دست است. مجموعه مقالات دو همایش پیشین زبان‌شناسی پیکره‌ای و مجموعه حاضر شاهدی بر این مدعا هستند.

امید است مجموعهٔ حاضر که دربرگیرندهٔ وجوه مختلفی از مطالعات زبان‌شناسی پیکره‌ای در موضوعات فرایندهای شکل‌گیری برخی پیکره‌های برچسب‌خورده و نخورده، پردازش زبان طبیعی با استفاده از پیکره‌های زبانی، کاربرد پیکره در آموزش زبان، تحلیل گفتمان و استخراج روال‌ها، الگوها و قواعد زبان فارسی بر اساس پیکره‌های موجود است، در بالندگی این حوزهٔ مطالعاتی مفید واقع شود.

لازم است از بزرگانی که در شکل‌گیری این همایش تلاش کردند، تشکر کنم؛ نخست از رئیس انجمن زبان‌شناسی ایران، سرکار خانم دکتر بلقیس روشن؛ مسئولان محترم پژوهشگاه علوم انسانی و مطالعات فرهنگی؛ رئیس

پژوهشکدهٔ زبان‌شناسی پژوهشگاه جناب آقای دکتر مصطفی عاصی؛ دبیر اجرایی سرکار خانم زهرا ابراهیم‌بانکی و همچنین شورای اجرایی همایش که بزرگوارانه در اجرای همایش بسیار کوشیدند. همچنین از سرکار خانم دکتر معصومه استاجی برای ویرایش چکیده‌های انگلیسی مقالات و از جناب آقای امیر احمدی مدیر محترم نشر نویسهٔ پارسی برای همراهی‌هایشان در چاپ مجموعهٔ مقالات سپاسگزارم.

بر خود لازم می‌دانم از کمیتهٔ علمی و هیأت داوری همایش و از استادان، دانشجویان و پژوهشگران گرانقدر که با ارسال مقاله خود به غنای علمی این همایش افزودند و دستاوردهای پژوهشی خود را سخاوت‌مندانه با علاقمندان این حوزه در میان گذاشتند صمیمانه سپاسگزاری کنم.

آزاده میرزائی

اردیبهشت ۱۳۹۸

تهیهٔ دادگان‌های گفتاری و متنی برای سامانهٔ بازشناسی خودکار مکالمات خلبان و واحدهای مراقبت پرواز

محمد بحرانی[۱]
مهسا آزادمنش[۲]

چکیده

این مقاله به مرحلهٔ آماده‌سازی پیکره‌های گفتاری و متنی برای یک سامانهٔ بازشناسی گفتار خاص‌منظوره به نام «سامانهٔ بازشناسی خودکار مکالمات خلبان و واحدهای مراقبت پرواز» می‌پردازد. در این مقاله، در اولین مرحله از طراحی سامانهٔ بازشناسی گفتار، به تهیهٔ دادگان‌های گفتاری و متنی موردنیاز برای آموزش سامانه می‌پردازیم. بدین‌منظور، بخش‌هایی از مکالمات واقعی صورت‌گرفته بین خلبان‌ها و برج مراقبت را از واحد مراقبت فرودگاه مهرآباد با اخذ مجوزهای لازم دریافت کردیم. مکالماتِ جمع‌آوری‌شده ابتدا پالایش شده و بخش‌های اضافی آن حذف می‌گردد و سپس توسط افراد خبره، متن معادل با آنها مطابق با یک سری استانداردهای خاص، تولید می‌شود. در مرحلهٔ بعد، فایل‌های صوتی مکالمات به همراه معادل متنی آنها به قطعات کوچکتر تقسیم می‌شوند؛ همچنین صورت‌های واجی انواع کلمات موجود در متن به صورت دستی تولید می‌شود. برای تهیهٔ دادگان متنی نیز علاوه بر متون مربوط به مکالمات، داده‌های متنی دیگری نیز از منابع مرتبط جمع‌آوری می‌شود. داده‌های متنی، مورد پالایش و یک‌دست‌سازی قرار می‌گیرند و درنهایت کلمات موجود در آنها استخراج می‌شود تا در مراحل بعد واج‌نویسی گردند. این کلمات به همراه کلمات استخراجی از دادگان صوتی، مجموعه واژگان سامانهٔ بازشناسی گفتار را تشکیل می‌دهند. در این پژوهش، در حدود ۱۵۲

[۱] استادیار، دانشکدهٔ علوم ریاضی و رایانه، دانشگاه علامه طباطبائی؛ bahrani@atu.ac.ir
[۲] دانش‌آموختهٔ کارشناسی ارشد زبان‌شناسی رایانشی، دانشگاه صنعتی شریف؛ mahsa.azadmanesh@yahoo.com

دقیقه مکالمهٔ صوتی به صورت تقطیع و برچسب‌دهی‌شده و همچنین یک دادگان متنی پالایش‌شده با حدود ۶۳۴۰۰ کلمه جمع‌آوری شده است. این داده‌ها در مراحل بعدی پژوهش، برای آموزش مدل‌های صوتی و زبانی سامانهٔ بازشناسی گفتار به کار می‌روند.

کلیدواژه‌ها: بازشناسی گفتار، دادگان مکالمات هوانوردی، پیکرهٔ خاص‌منظوره

۱. مقدمه

۱-۱. بیان مسئله

در حال حاضر در کشور بعد از بروز حوادث و رویدادهای مخاطره‌آمیز هوانوردی، مکالمات صورت‌گرفته بین خلبان‌ها و کنترل‌کننده‌های ترافیک هوایی (برج مراقبت)، توسط سازمان هواپیمایی کشوری مجدداً بررسی شده و این مکالمات توسط تیمی آشنا به مباحث هوانوردی، با صرف هزینه و وقت بسیار زیاد، به صورت دستی به متن تبدیل می‌شوند. این متون بعداً در صورت لزوم در مراجع قضایی مورد استفاده قرار می‌گیرند. «سامانهٔ بازشناسی خودکار مکالمات خلبان و واحدهای مراقبت پرواز» به پیاده‌سازی خودکار این مکالمات (تبدیل مکالمات به متن) کمک می‌کند تا در مواقع ضروری از متون مکالمات استفاده شود.

هدف نهایی ما در این پژوهش طراحی یک سامانهٔ بازشناسی گفتار است که کار متنی‌سازی را به‌صورت خودکار انجام دهد تا در وقت و هزینهٔ نیروی انسانی صرفه‌جویی شده و نیروی انسانی فقط نقش ناظر را داشته باشد.

ازآنجاکه برای آماده‌سازی هر سامانهٔ بازشناسی گفتار خاص‌منظوره (و با شرایط محیطی خاص)، ابتدا باید یک دادگان از سیگنال‌های گفتاری در همان زمینهٔ موردنظر و همچنین یک دادگان متنی از متونِ همان وظیفه[1] تهیه کرد، در این پژوهش و در اولین مرحله از طراحی سامانهٔ بازشناسی

[1] task

گفتار، به جمع‌آوری و پالایش مکالمات واقعی صورت‌گرفته بین خلبان‌ها و برج مراقبت می‌پردازیم و همچنین متن مکالمات را نیز به صورت دستی تهیه می‌کنیم. در مراحل بعدی پژوهش این داده‌ها برای آموزش مدل‌های صوتی و زبانی سامانهٔ بازشناسی گفتار متن‌باز کَلدی[1] (پُوی[2] و همکاران، ۲۰۱۱) به کار می‌روند تا در نهایت «سامانهٔ بازشناسی خودکار مکالمات خلبان و واحدهای مراقبت پرواز» طراحی و توسعه داده شود.

۱-۲. اهمیت و ضرورت انجام پژوهش

به دلایل زیر سامانه‌های معمول بازشناسی گفتار را نمی‌توان به منظور بازشناسی مکالمات خلبان و برج مراقبت به کار برد و باید با جمع‌آوری دادگان (گفتاری و متنی) مرتبط با این وظیفه، یک سامانهٔ بازشناسی گفتار مختص این کار طراحی کرد:

- به دلیل وجود نوفهٔ[3] زیاد در مکالماتِ هوایی، شرایط و محیط آکوستیکی مکالمات با محیط‌های آرام اداری که سامانه‌های معمول بازشناسی گفتار در آنها آموزش دیده‌اند، بسیار متفاوت است. بنابراین دادگان‌های گفتاری معمول که برای آموزش مدل‌های صوتی در سامانه‌های بازشناسی گفتار استفاده می‌شوند، در اینجا کارایی نداشته و باید از داده‌های واقعی مربوط به مکالمات هوایی استفاده کنیم تا بتوانیم گفتار نوفه‌ای را مدل‌سازی کنیم یا سامانه‌های موجود را با این داده‌ها تطبیق[4] دهیم.
- مکالمات هوانوردی دارای ادبیات خاص خود هستند و علاوه‌بر اینکه از قالب‌های خاصی پیروی می‌کنند، دارای کلمات و واژه‌های تخصصی

---

[1] Kaldi
[2] Povey
[3] noise
[4] adapt

خاصی هستند که در مکالمات روزمره به کار نمی‌روند. بنابراین مدل‌های زبانی موجود در سامانه‌های معمول بازشناسی به‌خوبی قادر به پیش‌بینی این سبک از جملات و واژه‌ها نیست. بنابراین باید مدل زبانی و مجموعه واژگان خاص مکالمات هوانوردی را فراهم کرد و آنها را در سامانهٔ بازشناسی به کار برد و یا اینکه مدل‌های زبانی موجود را با آنها تطبیق داد.

به علت حساسیت بالای هوانوردی، موضوع نظارت انسان بر روی تبدیل مکالمات به متن حذف نخواهد شد. ازآنجایی‌که در کشور ما تبدیل مکالمات به دادهٔ متنی زمانی صورت می‌گیرد که اتفاقی مخاطره‌آمیز یا سانحه‌ای رخ داده باشد و این داده‌های متنی در بسیاری از موارد در مراجع قضایی کاربرد دارند، نظارت انسان برروی پیاده‌سازی مکالمات، لازم بوده و سامانهٔ بازشناسی خودکار مکالمات نقش تسریع و تسهیل این روند را بر عهده خواهد داشت.

طبق استانداردهای بین‌المللی، مکالمات هوانوردی باید به زبان انگلیسی و با استفاده از اصطلاحات استاندارد هوانوردی صورت گیرد؛ بنابراین شاید این ایده به ذهن برسد که بتوان از سامانه‌های خارجی که بدین‌منظور طراحی شده‌اند استفاده کرد. ولی به دلایل زیر، سامانه‌های خارجی مناسب استفاده در داخل نیستند:

- گاهی خلبانان یا کنترل‌کننده‌های ترافیک هوایی، از تلفظ و اصطلاحات استاندارد پیروی نمی‌کنند و عموماً از زبان مادری خود (اکثراً فارسی) برای توصیف شرایط هوایی استفاده می‌کنند. بنابراین سامانه‌های خارجی نمی‌توانند پاسخگوی نیاز هوانوردی کشور باشند. از طرفی حتی در صورت صحبت به زبان انگلیسی نیز، لهجهٔ بیان مکالمات با لهجهٔ افراد بومی انگلیسی‌زبان (یا افرادی با زبان مادری غیرفارسی) بسیار متفاوت

است. بنابراین سامانه‌های بازشناسی گفتار که با لهجه‌های بومی انگلیسی یا سایر لهجه‌ها آموزش دیده‌اند، نمی‌توانند مناسب استفاده در داخل کشور باشند.

- بااینکه در تمام دنیا مکالمات هوانوردی دارای نوفهٔ زیاد هستند، ولی نوع نوفه بسته به کشورهای مختلف و کانال‌های مخابراتی مختلف بسیار متفاوت است. درنتیجه نمی‌توان سامانه‌ای که با یک نوفهٔ خاص تطبیق یافته را برای بازشناسی سیگنال‌های صوتی با نوفهٔ متفاوت به کار برد.
- در مکالمات هوانوردی کشور ما، نام شهرها و فرودگاه‌هایی به کار می‌رود که مختص کشور ایران بوده و درنتیجه در واژگان سامانه‌های خارجی وجود ندارد. بنابراین سامانه‌های بازشناسی خارجی قادر به تشخیص آنها نیستند.

مجموع دلایل بالا اهمیت و ضرورت انجام پژوهشی مستقل در داخل کشور را نشان می‌دهد.

۱-۳. چالش‌های تحقیق

عدم دسترسی به حجم زیادی از داده، وجود نوفه، وجود لهجه‌های مختلف، عدم رعایت اصطلاحات هوانوردی و صحبت به زبان مادری به جای انگلیسی، از چالش‌های کار بازشناسی گفتار هوانوردی در ایران است.

یکی از چالش‌های دست‌وپاگیر در روند انجام این پژوهش، عدم دسترسی به داده‌های هوانوردی بود. ازآنجاکه مکالمات هوانوردی عمدتاً محرمانه بوده و در مورد برخی از آنها ملاحظات امنیتی برقرار است، دسترسی به آنها بسیار مشکل بوده و نیازمند طی مراحل اداری پیچیده و اخذ مجوزهای لازم می‌باشد.

چالش دیگری که در کار وجود دارد رعایت نکردن استانداردهای هوانوردی در برخی از مکالمات است. همان‌طورکه گفتیم، مکالمات هوانوردی باید به زبان انگلیسی صورت گیرد؛ ولی گاهی خلبانان یا واحدهای کنترل ترافیک، این امر را رعایت نمی‌کنند. وجود زبان‌ها و گویش‌های مختلف، مانند فارسی، آذری، کردی و غیره در مکالمات انگلیسی هوانوردی از چالش‌های دیگر این پژوهش است.

۲. مرور مختصری بر پیشینه پژوهش

اولین سیستم تشخیص گفتار هوانوردی در جهان وُیس‌فلایت[1] نام دارد که توسط ادارهٔ هوانوردی فدرال آمریکا[2] برای استفاده در هواپیماهای غیرنظامی مجاز شمرده شده است (ویس‌فلایت، ۲۰۱۶). این سامانه از طریق کلید نصب‌شده برروی یوک[3] فعال می‌شود. خلبان، کلمات اختصاری مربوط به ایستگاه‌های بین‌راهی (که شامل سه حرف الفبای انگلیسی است) را می‌گوید و سامانه، این سرواژه‌ها را در واحد ناوبری وارد می‌کند.

اگر قرار باشد این سرواژه‌ها به‌صورت دستی وارد سامانه شوند خلبان باید برای هر جایگاه، پیچ مشخصی را آن‌قدر بچرخاند تا نمایشگر به حرف موردنظر برسد. سپس به جایگاه بعد بپرد و دوباره پیچ مربوط به حروف الفبا را بچرخاند تا جایگاه بعد نیز پر شود. انجام این کار بسیار زمان‌بر است. درحالی‌که با استفاده از سامانهٔ مذکور خلبان فقط با تلفظ حروف، کلمات اختصاری ایستگاه‌های بین‌راهی را وارد می‌کند. این سامانه فقط برای بازشناسی نام اختصاری فرودگاه‌های بین‌راهی از طریق بیان گفتاری حروف

---

[1] VoiceFlight

[2] Federal Aviation Agancy (FAA)

[3] سکان هدایت هواپیما

الفبا است (وارن[1]، ۲۰۱۶) و برای بازشناسی سایر مکالمات خلبان به کار نمی‌رود.

در پژوهش‌های بعدی، پاردو[2] و همکاران (۲۰۱۱) در دانشگاه پلی‌تکنیک مادرید، یک سامانهٔ بازشناسی گفتار هوانوردی طراحی کردند که قادر است مکالمات را همزمان از دو زبان اسپانیایی و انگلیسی تشخیص داده و به متن تبدیل کند. آنها در کار خود به دقت ۸۶٪ دست یافتند. ازآنجاکه تعداد کلمات و عبارات در مکالمات هوانوردی محدود است، تشخیص گفتار با دقت بالاتری صورت می‌گیرد.

جانسون[3] و همکاران (۲۰۱۳) نیز در ادارهٔ هوانوردی فدرال آمریکا و شرکت برین‌ونشنز[4]، سامانه‌ای به نام والس‌وُکس[5] ارائه دادند که قادر به تشخیص و درک مفاهیم موجود در مکالمات هوانوردی می‌باشد.

نگوین[6] و هولون[7] (۲۰۱۵) در مقالهٔ خود ضمن مروری بر سوابق موضوع، چالش‌ها و مشکلات سامانه‌های تشخیص گفتار هوانوردی را بررسی کرده و یک سیستم اولیه برای آن ارائه دادند و سپس نگوین در پایان‌نامه خود با به‌کارگیری اطلاعات زبانی، سامانه تشخیص مکالمات هوانوردی اولیه را بهبود داد (نگوین، ۲۰۱۶).

۳. جمع‌آوری و آماده‌سازی دادهٔ گفتاری

۳-۱. جمع‌آوری دادهٔ اولیه

[1] Warren
[2] Pardo
[3] Johnson
[4] Brain Ventions
[5] ValsVox
[6] Nguyen
[7] Holone

در اولین مرحله از کار تهیهٔ دادگان، نیازمند جمع‌آوری داده‌های خام هستیم. بدین‌منظور، طبق مذاکراتی با واحد مراقبت فرودگاه مهرآباد، مقداری دادهٔ صوتی ضبط‌شده از مکالمات هوانوردی را با اخذ مجوزهای لازم دریافت کردیم. حجم داده‌هایی که در اختیارمان قرار گرفت محدود بود و در بسیاری از موارد پیوستگی نداشت. بخشی از داده‌های صوتی مکالمات، برچسب متنی داشتند که البته به دلیل عدم تطابق کامل با فایل صوتی از این برچسب‌ها استفادهٔ چندانی نشد.

مکالماتی که در دست داریم شامل دو دسته هستند. دستهٔ اول مربوط به مکالمات روزمره خلبانان خطوط مختلف هواپیمایی و برج مراقبت و دستهٔ دوم، بخش‌هایی از ۹ رویداد جزئی پروازی که در سال ۱۳۹۱ برای هواپیماهای سه شرکت هواپیمایی هما، ماهان و نفت اتفاق افتاده‌اند. به دلیل تفاوت شرایط هوانوردی در مواقع عادی و در شرایط مخاطره‌آمیز، سبک جملات این دو دسته از داده‌ها متفاوتند. بخش‌هایی چند دقیقه‌ای از مکالمات دستهٔ دوم از روی نوار ضبط‌شده پیاده‌سازی شده و به متن تبدیل شده‌اند. به علت محرمانه بودن این داده‌ها امکان دسترسی به تمامی مکالمات این ۹ رویداد وجود نداشت. بنابراین بخش‌هایی غیرپیوسته از مکالمات بین کنترل‌کننده‌ها و خلبانان در اختیار ما قرار گرفت.

مکالمات دستهٔ اول یعنی مکالمات عادی روزمره شامل ۲۷۸ فایل صوتی به فرمت mp3 با حجم مجموعاً ۲۶۹ مگابایت و نرخ بیت[1] ۱۲۸ کیلوبیت بر ثانیه هستند. مکالمات تماماً به زبان انگلیسی می‌باشند.

مکالمات دستهٔ دوم خود شامل سه گروه از فایل‌های صوتی می‌باشد. گروه اول شامل ۱۲۳ فایل صوتی استریو به فرمت wav با حجم مجموعاً ۲۳۱ مگابایت دادهٔ خام می‌باشد که با نرخ نمونه‌برداری[2] ۸ کیلوهرتز (معادل نرخ

[1] Bit rate
[2] Sampling rate

بیت ۲۵۶ کیلوبیت بر ثانیه) ضبط شده‌اند. گروه دوم شامل ۱۰ فایل صوتی به فرمت wmv با حجم مجموعاً ۱۷۲ مگابایت و با نرخ بیت ۱۲۸ کیلوبیت بر ثانیه می‌باشد. گروه سوم نیز شامل ۱۱۳ فایل صوتی با فرمت mp3 با حجم مجموعاً ۵۹,۳ مگابایت و با نرخ بیت ۲۴ کیلوبیت بر ثانیه است. حدود هفتاد درصد از مکالمات دستهٔ دوم به زبان فارسی و حدود سی درصد به زبان انگلیسی صورت گرفته‌اند. در جدول ۱ جمع‌بندی این داده‌ها آمده است.

جدول ۱. ویژگی داده‌های صوتی خام (پیش از پالایش)

	۲۷۸	mp3	۱۲۸	۲۶۹
گروه اول	۱۲۳	wav	۲۵۶	۲۳۱
گروه دوم	۱۰	wmv	۱۲۸	۱۷۲
گروه سوم	۱۱۳	mp3	۲۴	۵۹,۳

در مکالمات به‌طور عمده از اصطلاحات هوانوردی استفاده شده است، درنتیجه ساختار جملات ساده و دایرهٔ لغات محدودند. نحو جملات ساده هستند. گوینده‌ها همه مرد هستند. میزان نوفهٔ داده‌ها خیلی زیاد است به‌طوری‌که فهم بیشتر مطالب برای کسی که گوشش با محیط پروازی آشنایی ندارد تقریباً غیرممکن است. در مکالمات گاهی نوفه به حدی زیاد می‌شود که بار اول خلبان یا برج مراقبت به‌طور درست متوجه پیام ردوبدل‌شده نمی‌شوند و مجبور هستند پیام را تکرار کنند.

۳-۲. پالایش داده‌های گفتاری

داده‌های گفتاری قبل از استفاده باید پالایش شوند. معادل متنی داده‌ها باید تولید شود. همان‌طورکه گفتیم، مکالمات هوانوردی باید انگلیسی باشند ولی بین آنها ممکن است مکالمات فارسی یا آذری نیز وجود داشته باشد. در این

پژوهش در مرحلهٔ اول تمرکز ما فقط برروی مکالمات انگلیسی است؛ بنابراین بخش‌های غیرانگلیسی در مرحلهٔ پالایش از سیگنال صوتی حذف می‌شوند. ضمناً داده‌ها باید از نظر میزان نوفه بررسی شوند و داده‌هایی که نوفه آنها از حدی بیشتر است حذف گردند. پالایش داده‌ها یکی از مراحل زمان‌بر و پرزحمت از کار تهیهٔ دادگان است زیرا بیشتر مراحل کار به صورت دستی انجام می‌شود. مراحل زیر به ترتیب برای پالایش داده‌ها به کار رفت:

۱. تبدیل فایل‌های گفتاری به فرمت wav: همان‌طورکه گفتیم، داده‌ها در سه فرمت wav، wmv و mp3 تحویل گرفته شدند. در اولین مرحله از پالایش داده‌ها، با استفاده از ابزار Cool Edit همهٔ داده‌ها به قالب wav تبدیل شدند.

۲. یکسان‌سازی نرخ نمونه‌برداری: در مرحلهٔ بعد، نرخ نمونه‌برداری همهٔ فایل‌های wav به حالت ۱۶ کیلوهرتز، ۱۶ بیت و مونو (معادل نرخ بیت ۲۵۶ کیلوبیت در ثانیه) درآمدند. همان‌طورکه گفتیم، داده‌های wav اولیه با نرخ ۸ کیلوهرتز نمونه‌برداری شده بودند. این داده‌ها در واقع upsample شدند تا به نرخ ۱۶ کیلوهرتز برسند. همچنین این فایل‌ها استریو بودند که به مونو تبدیل شدند. مرحلهٔ یکسان‌سازی نرخ نمونه‌برداری به صورت خودکار و با ابزار SoX انجام گرفت.

دلیل تبدیل داده‌ها به فرمت و نرخ نمونه‌برداری مذکور این است که ابزارهای بازشناسی گفتار که در مراحل بعد از ان استفاده می‌کنیم، عمدتاً با این فرمت کار می‌کند.

۳. حذف بخش‌های گفتاری فارسی: همان‌طورکه گفتیم بخش زیادی از داده‌ها در دستهٔ دوم، شامل مکالمات فارسی هستند. دلیل فارسی بودن این داده‌ها این است که معمولاً خلبانان در هنگام مواجهه با رویدادهای مخاطره‌آمیز به زبان مادری خود صحبت می‌کنند. بنابراین گفتار فارسی