কমপিউটারে স্বর চেনা

মুহসিন উদ্দীন আনওয়ার

কণ্ঠস্বর চেনা (voice recognition) অর্থাৎ কমপিউটার কিভাবে মানুষের উচ্চারিত শব্দকে 'শুনতে' ও 'বুঝতে' পারে সেজন্য বিভিন্ন পরীক্ষা নিরীক্ষা চলছে। ইতিমধ্যে মানুষ কমপিউটারকে দিয়ে কথা 'বলাতে' সক্ষম হলেও (যদিও তা নির্দিষ্ট প্যাটার্ণে উচ্চারিত হয় এবং গুটিকয়েক মুখস্ত বাক্যের ক্ষেত্রেই এখন পর্যন্ত তা সম্ভব হচ্ছে) কমপিউটার দিয়ে মানুষের স্বরকে বুঝতে পারা, অসংখ্য স্বর থেকে নিকটতম ও বিশিষ্ট স্বরকে আলাদা করতে পারা, error বা স্বর নয় এধরনের ধ্বনিকে আলাদা করতে পারা বেশ কঠিন এবং চ্যালেঞ্জপূর্ণ কাজ হয়ে দাঁড়িয়েছে।

মানুষের এইসব স্বাভাবিক কাজগুলো প্রকৃতি প্রদত্ত গুণ, সে তা ক্রমাগত শিখতে ও আয়ত্ত করতে পারে এবং পরিবেশের প্রতি তার সাড়া স্বতঃস্ফূর্ত। কমপিউটার কিন্তু তা নয়। কমপিউটার কিভাবে শিখবে (অর্থাৎ কিভাবে যুক্তি প্রয়োগ করে তার সিপিইউ মেমোরী ব্যবহার করে ডাটা প্রসেস ও সংরক্ষণ করবে) তা নির্ভর করে কমপিউটার নির্মাণকারীর উপর। কমপিউটার নির্মাতা ধীরে বা দ্রুত রান করে এরকম সিপিইউ তৈরী করতে পারে, বিশেষ হার্ডওয়্যার দিতে পারে যা সংবেদী অঙ্গকে প্রতিস্থাপিত করতে পারে এবং এসবের সমন্বয়ে যে কমপিউটার তৈরী হয়, তাকে যথাযথভাবে প্রোগ্রামিং করে লজিকটুকু তৈরী করেন প্রোগ্রামারগণ।

চিত্র-১ স্পাইক

এখন দেখা যাক, স্বর চিহ্নিতকরণে মানুষ (অর্থাৎ কমপিউটার প্রোগ্রামার ও নির্মাতারা) কিভাবে এগিয়ে গেছে। এ পর্যন্ত সংবেদী অঙ্গ কানকে প্রতিস্থাপিত করা হচ্ছে একটি শব্দ গ্রাহক যন্ত্রের মাধ্যমে যা শব্দ ও ধ্বনিকে ফ্রিকোয়েন্সীতে পরিণত করে মেমোরীতে রেখে দেবে, বা সিপিইউতে পাঠিয়ে দেবে পরবর্তী ধাপের কাজের জন্য।

এ হার্ডওয়্যারটি বিভিন্ন মানের হতে পারে। কোন কোন ভয়েস রিসিভার এমনভাবে প্রোগ্রাম করা থাকে, যাতে ত্রুটি বা নয়েজ কম হয়। নয়েজ কিভাবে কম করা যায়, তার উপর বিশেষ এলগরিদম বা যৌক্তিক পদ্ধতি প্রয়োগ করা হয়।

যেমন চিত্র-১এ একটি অনেকটা সর্পিল তরঙ্গ দেখা যাচ্ছে। যেখানে হঠাৎ একটি স্পাইক সৃষ্টি হয়েছে, সেটি স্বাভাবিকতা ভঙ্গ করছে অর্থাৎ অন্যান্য তরঙ্গ দশার ও বিস্তারের (Phase ও Amplitude) সাথে সঙ্গতিপূর্ণ নয়। সুতরাং এটি একটি নয়েজ।

সংবেদী হার্ডওয়্যারও তাই করে, নয়েজ সনাক্ত করে এবং তা দূর করে বিশেষ এলগরিদমের সাহায্যে। যদি সংবেদী হার্ডওয়্যারে তা 'রম'-এ প্রোগ্রাম করা না থাকে তবে এই প্রোগ্রামিংএর কাজটি প্রোগ্রামারকে সফটওয়্যারে করতে হবে।

এভাবে সংবেদী হার্ড-ওয়্যারটি নয়েজবিহীন তরঙ্গগুলো সংগ্রহ করবে। এখানে নয়েজ বলতে অন্য লোকের স্বরের মিশ্রণ বা কোন অনিয়মিত ও অনাকাঙ্খিত ধ্বনিকে বুঝানো হচ্ছে। এবং সংগৃহীত প্রতিটি তরঙ্গ পরবর্তী ধাপের প্রসেসের জন্য মেমরীতে বা ডাটা প্রবাহ পথে অবস্থান করবে।

এরপরে স্বভাবতঃই প্যাটার্ণ রিকগনিশন ও ম্যাচিং এর ব্যাপারটি এসে যায়। যদি দেখা যায় কোন তরঙ্গের ফ্রিকোয়েন্সীগুলোর সিরিজ পূর্বে মেমোরীতে থাকা বিশেষ ধ্বনির ফ্রিকোয়েন্সী সিরিজের সাথে মিলে যাচ্ছে তবে কমপিউটার 'বুঝতে' পারে যে, এটি ঐ ধ্বনি।

বিভিন্ন ধরনের ধ্বনির জন্য মেমোরীতে আলাদা আলাদা তথ্য ভান্ডার রাখতে হবে এবং যত সংখ্যক লোকের স্বর সংরক্ষণ করতে হবে। মেমরীর চাহিদাও তত বাড়বে। এতে প্রসেসিংয়ের সময়ও বাড়বে, তবে তার বিনিময়ে কর্মের ব্যাপকতাও বাড়বে।

অর্থাৎ কতটুকু বাড়বে তার একটি গাণিতিক প্রকাশ এমন– কর্মের ব্যাপকতা α প্রক্রিয়াকালীন সময় X প্রয়োজনীয় মেমরী।

যদি মেমোরী ষ্টোরেজ বা স্মৃতি ভান্ডার কমাতে চাওয়া হয়, অর্থাৎ শুধুমাত্র একটি ধ্বনির জন্য একটিই ফ্রিকোয়েন্সী সিরিজ থাকবে এবং বিভিন্ন লোকের ধ্বনির জন্য ঐ সিরিজকে একটি গাণিতিক সূত্র প্রয়োগ করে বিভিন্ন আকার প্রয়োগ করে তুলনা করা হয়, তবে আবার ডাটা প্রসেসিংয়ের সময়টা বেড়ে যাবে। এতে প্রোগ্রামিংয়ের জটিলতাও কিছুটা বেড়ে যাবে।

ধ্বনিকে তুলনা করার জন্য শুধু ফ্রিকোয়েন্সী জমা করাই যথেষ্ট নয়, সময়ের বিরতিও বেশ গুরুত্বপূর্ণ। সময়ের অল্প বিরতি শব্দের বৈশিষ্ট্য নির্দেশ করবে এবং সময়ের দীর্ঘ বিরতি নির্দেশ করবে বাক্যের সম্পূর্ণতা। এধরনের সময়ের বিরতি বিভিন্ন লোকের ক্ষেত্রে বিভিন্ন হবে। অর্থাৎ এখানেও ব্যাপকতা আনার জন্য বিশেষ গাণিতিক সূত্র তৈরী করতে হবে।

এইসব পদ্ধতির একটি বড় ধরনের দুর্বলতা হচ্ছে প্রতিবারে কোন ধ্বনিকে চেনার জন্য ও আলাদা করার জন্য জটিল প্রোগ্রাম ও সময় বাঁচাবার দরকার হয়।

বর্তমানে আন্তর্জাতিকভাবে যে পদ্ধতি অনুসরণ করা হয়, তা হলো প্রথমে কোন ব্যক্তির স্বরকে আলাদাভাবে রেকর্ড করে তার জন্য নির্দিষ্ট গাণিতিক সূত্রের ফেক্টর গণনা করা হয় ধারণকৃত কোন ষ্ট্যান্ডার্ড ধ্বনির ফ্রিকোয়েন্সী, এমপ্লিটিউড ও সিকোয়েন্স এর সাথে তুলনা করে।

পরবর্তীকালে যখন সেই একই ব্যক্তি কথা বলবে, তখন কমপিউটার 'অভিজ্ঞতা'র দ্বারা বুঝতে পারবে ঠিক কোন ব্যক্তি কথা বলছে এবং তার জন্য নির্দিষ্ট সূত্র তৈরী আছে কিনা। তখন যন্ত্রটি সব ধরনের সূত্র প্রয়োগ না করে বারে বারে ঐ নির্দিষ্ট সূত্রই প্রয়োগ করবে যাতে প্রসেসিংয়ের সময় কম লাগবে এবং স্বাভাবিক প্রোগ্রামেই কাজ হবে।

পূর্বেই যদি নির্দিষ্ট সূত্রটি তৈরী করা না থাকে তবে শুধুমাত্র শক্তিশালী ক্ষমতাসম্পন্ন প্রোগ্রাম ও কমপিউটারের পক্ষে ষ্ট্যান্ডার্ড ধ্বনি থেকে অপরিচিত ব্যক্তির 'ধ্বনি' বোঝা সম্ভব।

এখন পর্যন্ত খুবই সীমিত আকারে ধ্বনি চিহ্নিতকরণ করা যাচ্ছে, কারণ এমন কোন জেনারেলাইজড্ এলগরিদম তৈরী করা যাচ্ছে না যেটি পরিবর্তিত ধ্বনির সাথে যে কোন সময়ে ম্যাচ করতে সক্ষম হবে। অর্থাৎ যদি দু'জন বা ততোধিক লোক কথোপকথন করে, তখন পৃথক পৃথক সময়ে তা রেকর্ড ও বুঝতে সক্ষম হবে ঠিক কোন ব্যক্তি কথা বলছে এবং তার নির্দিষ্ট সূত্র প্রয়োগ করতে পারবে কিনা।

যেহেতু মানুষের কথা বলার ঢং ও ধ্বনির উচ্চারণভঙ্গি বিভিন্ন সময় পরিবেশ ও প্রয়োজনীয়তা অনুযায়ী বিভিন্ন হয়, সেজন্য কমপিউটারকে এনালাইসিস করতে হবে সবধরনের সমন্বয়ের উপর। এটি প্রকৃতই সময় সাপেক্ষ এবং জটিল প্রসেস।

একারণেই ভয়েজ রিকগনিশন এখন পর্যন্ত সীমিত এবং খুব কম ক্ষেত্রেই তা প্রযুক্ত হচ্ছে।

আশা করা যায় অদূর ভবিষ্যতে কৃত্রিম বুদ্ধিমত্তা প্রযুক্তি ব্যবহার করে কমপিউটারকে অভিজ্ঞতা ও সৃষ্টিশীলতা এ দুয়ের সমন্বয়ে এমন পর্যায়ে উন্নীত করা সম্ভব হবে যাতে মানুষ যেভাবে শিশুকাল থেকে আস্তে আস্তে স্বর চিনতে শিখে তেমনিভাবে কমপিউটার স্বর চিনতে পারবে।

স্বর চেনার ব্যাপারটি যখন বাস্তবভাবে সব স্থানে কার্যকরভাবে ব্যবহৃত হবে তখন পাল্টে যাবে বর্তমান কমপিউটারের গ্রহণ/নির্গম সিস্টেম। ডাটা কমিউনিকেশনে ভয়েস হলো সবচাইতে নমনীয় ডাটা যেটি ত্রুটিপূর্ণ হলেও উদ্ধার সহজ এবং ডাটা কমিউনিকেশন খুব সহজ এবং স্বল্পব্যয়সাধ্য। আবার মানুষের সাথে তথ্য আদান-প্রদানে ভয়েস হলো সবচাইতে ভালো মিডিয়া।

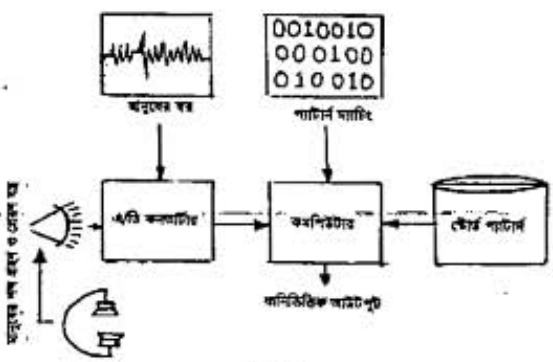
অন্য একটি প্রযুক্তি হলো– স্পিকিং হার্ডওয়্যার এবং সফটওয়্যার। দুটোই স্বর চিহ্নিতকরণের বিপরীত প্রক্রিয়া মাত্র।

ভয়েস ট্রান্সমিশন, রিকগনিশন, প্রসেসিং এবং আউটপুটের ক্ষেত্রে পাশ্চাত্যের ও জাপানের অগ্রসর প্রযুক্তির প্রয়োগ এখনই বেশ লক্ষণীয়। বিশেষ বিশেষ ব্যক্তির ধ্বনি শুনে কমপিউটার তা সনাক্তকরণ করতে পারে এবং উত্তর দিতে পারে ব্যক্তির আকাঙ্খিত অনুসন্ধান অনুযায়ী।

এটাই সাফল্য। স্বর সনাক্তকরণ প্রক্রিয়া কমিয়ে দেবে আমাদের এবং কমপিউটার উভয়েরই কাজে এবং সময়ের অপচয়। তথ্য আদান প্রদান সম্ভব হবে আরো দূরদূরান্তব্যাপী। এবং সম্ভব হবে ভিন্ন গ্রহ থেকে আগত সংকেত প্রসেস করে ভিন্নগ্রহবাসী কোন প্রাণী যদি থেকে থাকে তাদের কণ্ঠস্বর ধারণ করা এবং প্রোগ্রামিং এর মাধ্যমে প্রসেস করে বুঝতে পারা তারা কি উচ্চারণ করছে এবং তার অর্থই বা কি।

বিভিন্ন ভাষাভাষীদের জন্য কথোপকথন হবে অত্যন্ত সাধারণ। ভয়েস রিসিভারে তারা কথা বলবে। প্রোগ্রামকৃত কথা যখন আউটপুট আকারে বের হবে, তা সম্পূর্ণ ভিন্ন ভাষায় অপর ব্যক্তির নিকট প্রকাশ পাবে।

এভাবেই কমপিউটার প্রযুক্তির কল্যাণে একদার স্বপ্ন বাস্তবে পরিণত হবে।



চিত্র-২

নোট ঃ

কিভাবে নয়েজ বিহীন ধ্বনি রেকর্ড করা হয়

প্রিপ্রসেসিং এবং মাইক্রোফোন প্রযুক্তিতে নয়েজ ভালোভাবে কমানো যায়। প্রাপ্ত ধ্বনিকে ডিজিটাল ফিল্টার দিয়ে বিচ্ছিন্ন ধ্বনিসমষ্টি রূপান্তর করা হয় এবং এগুলোকে বলা হয় টোকেন বা প্যাকেট।

কতগুলো নিয়মের সেট প্রয়োগ করে টোকেনগুলোকে তুলনা করা হয় আদর্শ স্বরের ধরনের সাথে। যখন নিয়মের শর্তগুলো খাপ খেয়ে যায়, তখন কমপিউটার বুঝতে পারে যে, এটি একটি আকাঙ্খিত ধ্বনি। ধ্বনিসমষ্টির একটি সিরিজকে শব্দের প্যাটার্ণ এর সাথে তুলনা করে নির্দিষ্ট উচ্চারিত শব্দটি বাছাই করা হয়।

এসব তুলনাকরণের বিভিন্ন এলগরিদম রয়েছে। যেমন Hidden Markov Modelling (HMM) হলো সবচাইতে বেশী ব্যবহৃত এলগরিদম যেখানে একটি নির্দিষ্ট টোকেনের শক্তিকে একটি মডেল সিরিজের সাথে মিললে সেটিকে রিসিভ করা হয়। এভাবে প্রাপ্ত মডেলগুলোর সমষ্টি মিলে তৈরী হয় একটি নির্ধারিত ধ্বনি।

আরেকটি এলগরিদম হলো Spectral Peak Picking (SPP) যেখানে প্রত্যেকটি প্যাকেট দুটো তেলু এসাইন করে যার মাধ্যমে নির্দিষ্ট ফ্রিকোয়েন্সী রেঞ্জে সর্বোচ্চ পিক নির্ধারণ করা হয়। এই পিকগুলো আবার পাঠানো হয় একটা ফিল্টার ব্যাংকে এবং যেটি ফিল্টারকে অতিক্রম করতে পারে তা রেকর্ড করা হয়। একটি সম্ভাব্যতা ফাংশন এর সাহায্যে নির্ধারণ করা হয় ঠিক কোন ধ্বনিটি উচ্চারণ করা হয়েছিল।

এভাবে ধ্বনি থেকে প্রাপ্ত শব্দগুলোকে অর্থবহ কিনা তা পরীক্ষা করে দেখা হয়।

প্রচলিত এপ্লিকেশন প্যাকেজসমূহ ঃ

(১) প্রম্পট, লিসেন, লগপ্যাকেট ঃ

মার্কনি (টকম্যান) ঃ

এটি বহুযোগ্য ডাটা টার্মিনাল এতে মাইক্রোফোন সংযুক্ত একটি হেডসেট থাকে যার সাথে কমপিউটারে অপটিক্যাল কমিউনিকেশন থাকে। ধ্বনি সনাক্তকৃত হয়ে কমপিউটারে প্রেরিত হয়। এটি ২৫০টি বিভিন্ন শব্দ সনাক্ত করতে পারে যা আগে থেকে ষ্টোর করতে হয়। এর অন্যতম বৈশিষ্ট্য হলো এটি একাধিক ব্যক্তির উচ্চারিত শব্দ সনাক্ত করতে পারে।

লোগোস (লোজিকা) ঃ

এটি অনেকটা মার্কনি টকম্যান সিস্টেমের মতই। পূর্ব থেকে শব্দ ষ্টোর করতে হয়। ৮০২৮৬ মেশিনের সাথে একটি ফিল্টার সার্কিট লাগানো হয়, যার সাথে টিএমএস ৩২২০ সংযুক্ত থাকে। একটি স্পীচ বাফারে ৫.২০ সে. দৈর্ঘ্যে ম্যাসেজ ষ্টোর করা যায়। সক্রিয় শব্দ বিভিন্ন জটিলতা অনুযায়ী ২০-২৪০টি পর্যন্ত হতে পারে।

ট্যানগোরা (আইবিএম)

এটি একটি ভয়েস টাইপরাইটার। ২০,০০০ পর্যন্ত শব্দ সনাক্ত করতে পারে। শব্দগুলো আলাদা করে উচ্চারণ করতে হয়। ভিন্ন ভিন্ন ব্যক্তির জন্য বিশিষ্ট শব্দ প্যাটার্ণগুলো আগে থেকে ষ্টোর করতে হয়। প্রতি ১০ মিলি সেকেন্ড পরপর স্পেকট্রাল প্যাটার্ন মিলিয়ে দেখা হয় ২০টি ভিন্ন ভিন্ন ফিচারের সাথে। অতিরিক্ত চিপ বোর্ড সংযুক্ত থাকে।

ভিকশন সিস্টেমস (ড্রাগন কার্ডওয়েল এপ্লাইড ইনটেলিজেন্স) ঃ

প্রাথমিকভাবে ৪০০ প্রতিনিধিত্বশীল শব্দ প্রয়োজন। প্রত্যেকটি শব্দের পরে বিরতি দিতে হয়। কোন শব্দ পরিষ্কারভাবে বুঝতে না পারলে কমপিউটার ২/৩টি শব্দকে পছন্দ করার জন্য প্রম্পট দেয়। মেডিক্যাল ডিকটেশনে ব্যবহৃত হয়। ১০,০০০-৩০,০০০ শব্দ সনাক্ত করতে পারে।

ম্যাক্রোশিফ্ট (মার্কনি) ঃ

ডাটা স্থানান্তরে ব্যবহৃত হয়। স্থানান্তর হার প্রতিসেকেন্ডে ১২৫ কিলোবাইট। ব্যাংক এবং শিল্প নিরীক্ষণে এর প্রয়োগ সম্ভব। শব্দ ভান্ডার ৬৪০ শব্দের এবং একটানা ২০৫ সেকেন্ড ধ্বনি ষ্টোর করে। এতে একটি ভূয়াল আরএস ২৩২ সি লাগানো হয়।

(২) প্রম্পট, লিসেন, এ্যাপ্ট প্যাকেজ ঃ

এআরএস ১০০০ (মার্কনি) ঃ

এয়ারবর্ণ স্পিচ রিকগনাইজার, শব্দভান্ডারে রয়েছে ১০০ শব্দ। একইসাথে ব্যবহারকারীর সংখ্যা ২। ব্যাকগ্রাউন্ড নয়েজের কারণে উচ্চস্বরে উচ্চারিত কথাকে ঠিকমত রিকগনাইজ করতে পারে। মৌখিক নির্দেশকে হস্তচালিত নির্দেশের বিকল্প হিসেবে ব্যবহার করা যায়।

আরমাডা (রয়েল সিগনালস এন্ড রাডার এসটাব্লিশমেন্ট)

আরমাডা, HMM এর উন্নত স্ট্যাটিসটিক্যাল কৌশল ব্যবহার করে তৈরী শব্দ ভান্ডারে রয়েছে ৫০০ শব্দ।

স্পিচ ভিউয়ার (আইবিএম) ঃ

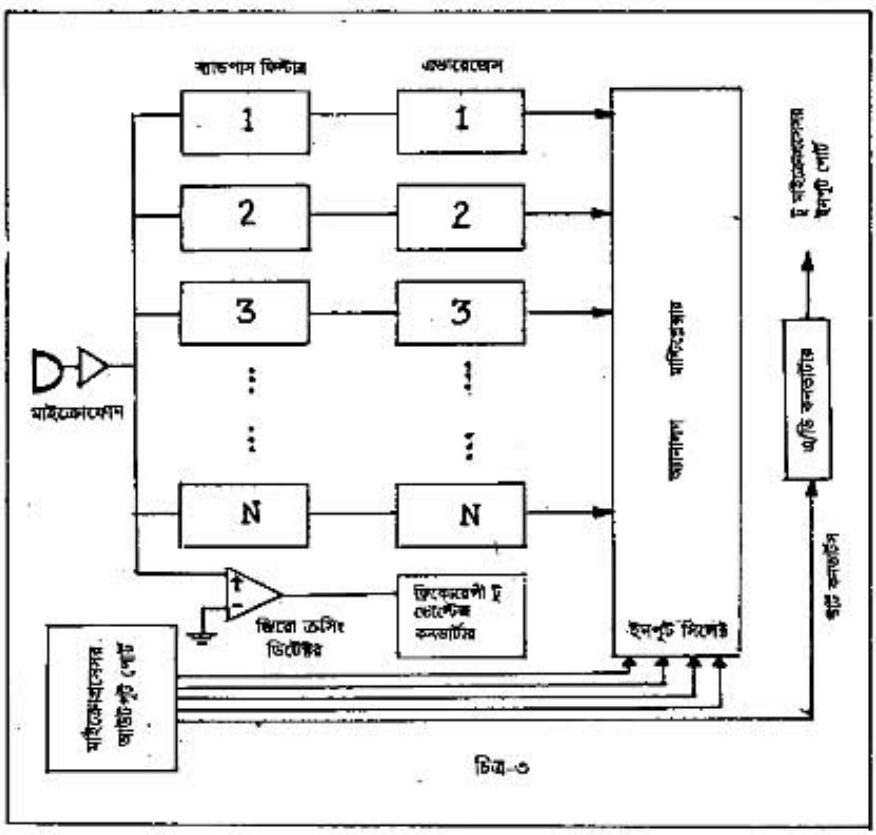
স্পিচ থেরাপি এবং বধির প্রশিক্ষণের কাজে ব্যবহৃত হয় স্পিচ ভিউয়ার। স্পিচ কে সরাসরি গ্রাফিক্সস ডিসপ্লে তে রূপান্তর করা হয় যেটি প্লে ব্যাক-এর মাধ্যমে সিনক্রোনাইজ করা যায়। ৬৪০ কিলোবাইট র‍্যাম পুরোপুরি ব্যবহার করে।

ভোডিস (লোজিকা, ব্রিটিশ টেলিকম, ক্যামব্রিজ বিশ্ববিদ্যালয়)

এটি টেলিফোন রেলওয়ে টাইমটেবল হিসেবে ব্যবহৃত হয়। প্রশ্ন ও উত্তর প্রক্রিয়ায় একজন প্রশ্নকারীর সঠিক লক্ষ্য জেনে নিয়ে উত্তর সরবরাহ করে। এভাবে সংক্ষিপ্ত সময়ে রেলওয়ে টাইমটেবল-এর টেলিফোন সার্ভিস দিতে পারে।

সানডিয়াল (লোজিকা এন্ড স্পিরিট)

এ প্রোগ্রামটি কথা বুঝতে পারে এবং বলতে পারে। ফ্রান্স, জার্মানি, ইটালী এবং ইংল্যান্ড-এর বিশেষজ্ঞরা যৌথভাবে এটা তৈরী করেছেন। টেলিফোন, ব্যাংকিং এবং হোটেল ও ট্রাভেল বুকিং-এ এটি ব্যবহৃত হয়। এটি HMM পদ্ধতির উপর ভিত্তি করে তৈরী। শব্দভান্ডারে রয়েছে ১০,০০০ শব্দ। চারটি ভাষাতেই প্রয়োজ্য।



চিত্র-৩

ধ্বনি ফিল্টারকরণ এবং কনভার্সন

সূত্র ঃ

1. Computer Chronicle
2. Microprocessor & Interfacing —By Dauglus V. Hall.
3. Gazette Chips —1990, February.
4. Bits on Computer —1990, April.
5. Syntactic Pattern Recognition — By K.S. Fu.

দ্রুত কমপিউটার জগৎ পেতে হলে

পাঠক সেবার জন্য ঢাকার নিম্নলিখিত কয়েকটি যায়গায় 'কমপিউটার জগৎ' বের হওয়ার কয়েক ঘন্টার মধ্যে পাওয়া যায় ঃ

মোস্তফা বুক ষ্টল-কলাবাগান বাস স্ট্যান্ড; জিনাত বুক সাপ্লাই লি ঃ- নিউ মার্কেট অনুপম জ্ঞান ভান্ডার-ঢাকা স্টেডিয়াম (দোতলা), সাগর পাবলিশার্স-নিউ বেইলী রোড, সুজনী-কমলাপুর রেলষ্টেশন ; মনা নিউজ কর্নার- পিজি হাসপাতাল শাহাবাগ, ঢাকা।

গ্রাহক হওয়ার জন্য বার্ষিক (রেজিষ্ট্রি ডাকে) দুইশত দশ টাকা ষান্মাসিক একশত দশ টাকা নগদ, মানি অর্ডার, চেক, ব্যাংক ড্রাফট-এ "কমপিউটার জগৎ" নামে ১৪৬/১, আজিমপুর রোড, ঢাকা-১২০৫ এই ঠিকানায় পাঠাতে হবে।।